

Concurrent Information Communication in Voice-based Interaction

By

Muhammad Abu ul Fazal

Supervised By

Dr. Sam Ferguson

Dr. Andrew Johnston

A Thesis Submitted in Fulfilment of the

DOCTOR OF PHILOSOPHY

School of Computer Science

Faculty of Engineering and Information Technology

University of Technology Sydney

July, 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Muhammad Abu ul Fazal declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in Information Technology, in the School of Computer Science, Faculty of Engineering and IT at the University of Technology Sydney.

This thesis is wholly my own work, unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 23/07/2019

Acknowledgment

Firstly, I would like to express my sincere gratitude to my supervisors Dr Sam Ferguson and Dr Andrew Johnston. From my initial offer to be a visiting research student at the University of Technology Sydney (UTS) to the offer of completing a PhD from UTS with fully funded Scholarship, I am truly grateful for the opportunities. Each of my supervisors helped, supported, and guided me through my research at UTS in an exceptional and unforgettable manner. I shall always remain thankful to Dr Ferguson and Dr Johnston.

Dr Ferguson's insights taught me how to address a research problem scientifically, how to conduct standardised experiments, how significant it is to apply statistical tests on results, how to present results visually and also how to make an academic argument. These are but a few research skills from a, what seems, unending list that I learned from Dr Ferguson.

I am also thankful to the Faculty of the Department of Computer Science, Quaid-i-Azam University (QAU), Islamabad, Pakistan for offering me admission in PhD. I am thankful to my former research supervisor Dr Shuaib Karim for supervising me in QAU and guiding towards international opportunities. I am also thankful to the Higher Education Commission of Pakistan for awarding me an IRSIP scholarship to support my studies at UTS. I wish there were an official agreement of researching collaborations between the QAU and UTS, which would have enabled me to continue my PhD as a joint degree program from both the universities together.

I am thankful to the office staff at the School of computer science, UTS, in particular, Margot, Janet, Teraesa and Reshma. All team members I met within

the School, provided a conducive research environment. I am also thankful to the staff at Graduate Research School at UTS, in particular, Jing and Grandia, who always remained welcoming of my questions and queries, resolving my problems in a timely manner. I also pay thanks to the office staff at CS department at QAU for the support they provided me.

Regards

Muhammad. Abu ul Fazal

Dedication

My thesis is dedicated to the people who have supported my goals, inspired me and challenged me academically to make it to this day.

Reflecting on my path that led me to this day, after completing matriculation, circumstances forced me to abandon studies for what I thought would be a permanent arrangement. I provided support to my family and assisted my father run a medical store. Happenstance led a past teacher to the medical store and I spoke to him about resuming studies. He suggested I join his evening academy to prepare for 12th-grade exams while still helping my father. I joined the academy.

I fondly remember one particular day at the academy when I inadvertently mentioned my name and my aspiration to my childhood friend Adnan; “Hello, I am Professor Doctor Muhammad Abu ul Fazal”. He asked *will you be?* I replied *maybe!* The title combined with my name pleasantly haunted my mind.

In 2003 when I was in my 3rd year of completing the Bachelor of Science, one of my teachers suggested that I extend my degree and complete the Master of Science so that I could later enrol in a PhD. Today, I submit my Doctoral thesis at the University of Technology Sydney. My earlier ‘maybe’ has become a ‘yes’.

For this day today, first and foremost credit goes to my mother, who during a severe financial crisis when I asked about my possibility to study a Bachelors, she replied: “Fazal, take admission and do it”. I saw her working days and nights taking on additional teaching activities to manage my education fees.

Thank you, Ammi. Thank you Great Abba G, you have always been an inspiration to me and supported a lot from the background. Thank you to my sisters, Saima Umar, Nafisa Umar, Shabiah Umar, Asma Junaid and my brothers

Junaid Umar, Safi Ullah, Akmal Ata, Waseem Asif, and lovely nieces and nephews for always remaining a great source of support and encouragement. Ghulam Mustafa, my dear friend, thanks to you too.

When I started my PhD my immediate family comprised of my wife Hadiah, our daughter Mahrukh and me. In the middle of my PhD, my son Sherdil joined our family. Without the support of Hadiah, I would have never been able to complete my doctorate studies abroad. Mahrukh, and Sherdil, I owe you a lot. The time I spent in Australia doing my PhD, you in your childhood living in Pakistan - you were the real owners of this time. Hadiah, I am eternally thankful to you. Thank you to my great father-in-law and mother-in-law and as well as siblings of my wife who offered well wishes for my studies and prayed for my success.

Abstract

Speech-based information is primarily communicated to users sequentially; however, users are capable of obtaining information from multiple sources concurrently. This fact implies that the sequential approach is under-utilising human perception capabilities and restricting users to perform optimally. In this research, two informal studies and two experiments were carried out for investigating concurrent communication of multiple voice-based information streams. The informal studies were carried out to understand users' interest and expectations in concurrent information communication and to examine whether users can comprehend concurrent information. In the first experiment, different designs for speech-based multiple information communication and the depth of comprehension by users in each design were tested. In the second experiment, various combinations of information streams presented concurrently and their viability regarding cognitive load were tested.

The results of the first study manifested user's interest in concurrent information communication design and supported the argument that users are able to discriminate and understand the concurrent voice streams using their selection and attention abilities. The results led to the second study, where users, including visually challenged users, expressed their expectations from such system and shared how would they prefer to interact with the systems providing concurrent information communication. Based on user's feedback, a web-based '*Vinfomize framework*' is designed to allow for concurrent communication of multiple information streams to users. Findings from the third study showed that concurrent speech-based information designs, involving intermittent form and a spatial dif-

ference in sources of the streams, provide satisfying comprehension of the content. The study further showed that users could comprehend both the main information and the detailed information. The fourth study showed that the perceived cognitive workload for the listening task in baseline condition and concurrent combinations remain the same; however, users response in preference and frequently using different combinations remain significantly lower than the baseline condition. The fourth study also showed that the combinations created with music were preferred the most by the users in concurrent combinations, followed by the song. From the information types providing speech-based information (non-music/song), result shows the intermittent form of communication creates the low cognitive workload in voice-based information communication.

Our research findings contribute to providing improvements in methods to communicate voice-based information efficiently under a large variety of application fields.

One Page Thesis

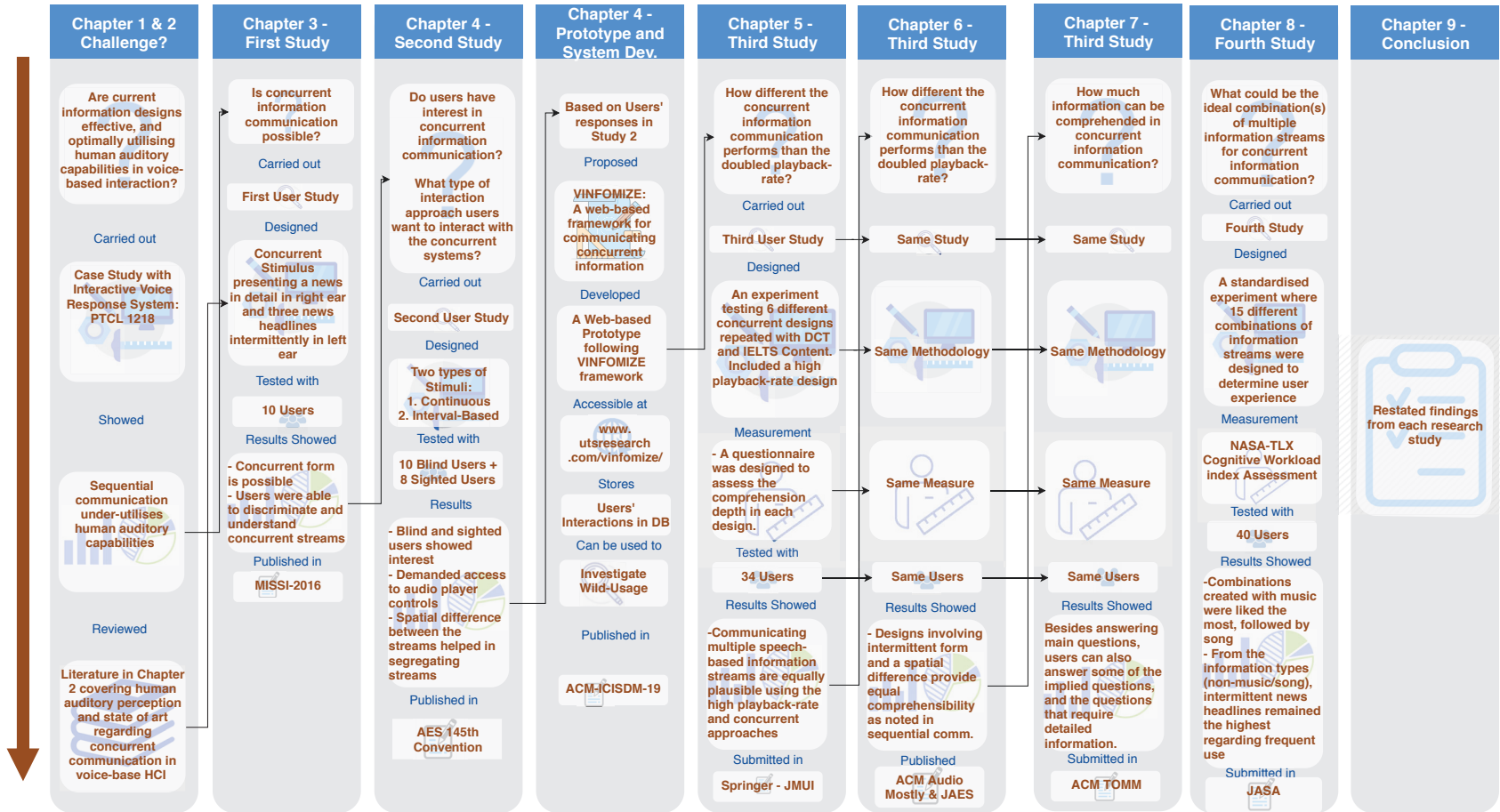


Figure 1 : **Research Overview:** Involved 102 users in total including visually challenged users in 4 different user studies, recorded 24000 user responses, and processed them to come up with the findings mentioned in this thesis.

List of Publications

1. M. A. u. Fazal and M. Shuaib Karim, "Multiple information communication in voice-based interaction," in *Advances in Intelligent Systems and Computing*. Springer, pp. 101–111
2. M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, "Concurrent Voice-Based Multiple Information Communication: A Study Report of Profile-Based Users' Interaction," in *145th Convention of the Audio Engineering Society*. Audio Engineering Society, 2018
3. M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, "Vinfomize: A framework for multiple voice-based information communication," in *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*. ACM, 2019, pp. 143–147
4. —, "Investigating Efficient Speech-based Information Communication - A Comparison between the High-rate and the Concurrent Playback Designs," *Journal on Multimodal User Interfaces (JMUI)*, vol. -, no. -, pp. 1–8, 2019, submitted
5. M. A. u. Fazal, S. Ferguson, and A. Johnston, "Investigating Concurrent Speech-based Designs for Information Communication," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ACM. New York, NY, USA: ACM, 2018, pp. 1–8
6. —, "Investigating Concurrent Speech-based Designs for Efficient Information Communication - Extended Analysis," *Journal of the Audio Engineering Society (JAES)*, vol. -, no. -, pp. 1–8, 2019, submitted

7. M. A. u. Fazal, S. Ferguson, and A. Johnston, "Evaluation of Information Comprehension in Speech-based Designs for Concurrent Audio Streams," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. -, no. -, pp. 1–18, 2018, submitted
8. —, "Investigating cognitive workload in concurrent speech-based information communication," *The Journal of the Acoustical Society of America (JASA)*, vol. -, no. -, pp. 1–20, 2019, submitted
9. A. Hussain, M. A. u. Fazal, and M. S. Karim, "Intra-domain user model for content adaptation," in *Smart Innovation, Systems and Technologies*. Springer, 2015, pp. 285–295

Contents

Publications	x
List of Tables	xvii
List of Tables	xvii
List of Figures	xviii
List of Figures	xviii
1 Introduction	1
1.1 Voice-based Interaction	1
1.2 Interactive Voice Response System: PTCL Case-Study	3
1.3 Human Abilities of Listening to Multiple Information Simultaneously	6
1.4 Multiple Information Communication in Voice-based Interaction . .	8
1.5 Motivating Scenarios	10
1.6 Research Aim	12
1.6.1 Research Questions	12
2 Literature Review	14
2.1 Human Auditory Perception	14
2.2 Psychological Studies Exploring Concurrency	22
2.3 Contemporary Research Studies on Concurrent Speech Interface . .	29
2.4 Summary of the Angles Reviewed	36
2.5 Our Research	36

3	Viability of Concurrent Information Communication	38
3.1	Aims & Motivation	38
3.2	Methodology	39
3.2.1	Stimulus Material	39
3.2.2	Design	39
3.2.3	Participants	40
3.2.4	Questionnaire	41
3.3	Results	41
3.4	Discussion	46
3.5	Limitations & Future Work	47
4	Viable Interaction Approach to Interact with the System Communicating Concurrent Information?	49
4.1	Aims & Motivation	50
4.2	Investigation	50
4.2.1	Participants	51
4.2.2	Study 2-A - Continuous: Stimulus & Questionnaire	52
4.2.3	Study 2-B - Continuous and Intermittent: Stimulus & Questionnaire:	54
4.2.4	Protocol	55
4.3	Results & Analysis	57
4.3.1	Qualitative Analysis	57
4.3.2	Quantitative Analysis	59
4.4	Discussion	60
4.5	Vinformize Framework	62
4.6	Web-based System Development based on the proposed Framework	67
4.7	Limitations & Future Work	68

5 A Comparison between High Playback-rate and Concurrent Design	71
5.1 Aims & Motivation	72
5.2 Method	73
5.2.1 Participants	73
5.2.2 Design	73
5.2.3 Material	77
5.2.4 Stimuli Information	78
5.2.5 Measures	79
5.2.6 Questionnaire	79
5.2.7 Apparatus	80
5.2.8 General Procedure	81
5.3 Results	81
5.3.1 Proportion Analysis	82
5.3.2 Comprehension Performance Analysis	83
5.3.3 Comprehension Depth Analysis	85
5.4 Discussion	87
5.5 Limitations and Future Work	88
6 Investigating Concurrent Speech-based Designs for Information Communication	90
6.1 Aims & Motivation	90
6.2 Methodology	91
6.3 Results	91
6.3.1 Concurrent Designs Analysis	92
6.3.2 Intermittent Designs in Detail	94
6.4 Discussion	98
6.5 Limitations & Future Work	99

7 Evaluation of Information Comprehension Depth in Speech-	
based Concurrent Designs	102
7.1 Aims & Motivation	102
7.2 Method	103
7.3 Results	103
7.3.1 Overall Comprehension Comparison between Streams . . .	104
7.3.2 Comprehension Depth for Concurrent Condition	106
7.3.3 Users' Experience	108
7.4 Discussion	109
7.5 Limitations and Future Work	113
8 Evaluating Various Combinations of Information Streams	
in Concurrent Information Communication	115
8.1 Aims & Motivation	115
8.2 Method	116
8.2.1 Participants	116
8.2.2 Design	117
8.2.3 Material	119
8.2.4 Stimuli Information	119
8.2.5 Measures	120
8.2.6 Apparatus	121
8.2.7 General Procedure	122
8.3 Results	122
8.3.1 Baseline vs. Concurrent	123
8.3.2 Concurrent Combinations	125
8.3.3 Information Streams Impact in Concurrent Communication	133
8.3.4 Impact of Presentation in Left — Right Ears	135
8.4 Discussion	138

8.5 Limitations and Future Work	143
9 Conclusion	145
Bibliography	149
Appendix A A Letter to the Institute Requesting the Participa- tion of Visually Challenged Persons in Study II	176
Appendix B Vinformize-based System Interface	178
Appendix C Ethics Application for Study III	180
Appendix D Interface with Selected Playable Audio Files URLs, and Questionnaire for Study III	190
Appendix E Ethics Application for Study IV	209
Appendix F Interface with Selected Playable Audio Files URLs, and Questionnaire for Study IV	215
Appendix G Publication 1	221
Appendix H Publication 2	232
Appendix I Publication 3	238
Appendix J Publication 4 [Submitted]	246
Appendix K Publication 5	257

Appendix L Publication 6 [Submitted]	266
Appendix M Publication 7 [Submitted]	276
Appendix N Publication 8 [Submitted]	295
Appendix O Publication 9	346

List of Tables

3.1	Participants Demography	40
3.2	Questionnaire & Users Responses	42
4.1	Users' Profiles	51
4.2	Profile-based User Groups	52
4.3	Questionnaire Study 2-A	54
4.4	Questionnaire Study 2-B:	56
5.1	Speech-based Concurrent Communication Designs	75
5.2	Results of the One-to-one Proportion Comparison	86
7.1	Proportion Comparison Test:	105
7.2	One-to-one Proportion Comparison Test Results	108
8.1	Combinations of Different Types	117
8.2	<i>Post hoc</i> Tukey HSD Analysis Comparing Concurrent Scales scales .	125
8.3	<i>Post hoc</i> Tukey HSD Analysis Comparing Speech-based Concurrent Combinations	128
8.4	<i>Post hoc</i> Tukey HSD Analysis Comparing Music-based Concurrent Combinations	131
8.5	<i>Post hoc</i> Tukey HSD Analysis Comparing Stream Types	133

List of Figures

1	Research Overview	viii
1.1	Voice-based Interaction	2
1.2	IVR Case Study	4
1.3	Auditory Perceptual Dimensions	7
1.4	Auditory System Mechanisms and Processes	8
1.5	Overlay Example	9
2.1	Schematic View of the Periphery Auditory System	15
3.1	Distinguish Secondary Voice	44
3.2	Interesting News Topic	45
3.3	Multiple Information Preference	45
4.1	Continuous Stimulus Design for Study 2-A	53
4.2	Continuous and Intermittent Stimulus Design for Study 2-B	53
4.3	Group-wise Responses in Both Studies	59
4.4	Group-wise Responses in Basic and Advanced Questions	60
4.5	All Users' Responses in Both Studies	61
4.6	Vinfomize Framework	63
5.1	Stimuli Designs	76
5.2	The Proportion of User Responses	83

5.3	Percentage of Correct Answers	84
5.4	Comprehension Depth	86
6.1	Proportion of Users Responses	92
6.2	Percentage of Correct Answers	94
6.3	Competing and Non-competing Questions	95
6.4	Percentage of Correct Answers	96
6.5	Deep Intermittent Analysis	97
7.1	Percentage of Correct Answers	104
7.2	Users' Comprehension in each Stream w.r.t MIS, MII, DTS, DTI . .	107
7.3	Participants Preference	109
8.1	Concurrent Stimulus Design	118
8.2	User Experience in Baseline Condition	123
8.3	Users' Experience in Concurrent Communication	125
8.4	Users' Experience in each Combination	126
8.5	Users' Experience regarding each Information Type	134
8.6	Impact of Users' Experience (Stress)	137
8.7	Impact of Users' Experience (Acceptance)	137
8.8	Perceived Workload Index Score	138
8.9	Ratings for Frequent and Like Scales	139
8.10	Order of Combinations	140
8.11	Order of Information Types	141

Chapter 1

Introduction

The focus of this research is to investigate the possibilities of communicating multiple speech-based information streams concurrently. Presently, in an interaction method, the system communicates information to a user sequentially whereas users are capable of noticing, listening and comprehending multiple voices simultaneously. In this research, the possibilities of communicating information concurrently to a user are explored, and various speech-based concurrent designs are tested to identify whether efficient concurrent information can be adequately comprehended.

In this introductory chapter, after discussing the notion of voice-based interaction the under-utilisation of the human auditory capabilities in communicating information to the users by the system will be highlighted and supported with the help of an interactive voice response system (IVR) case study. Human auditory capabilities are briefly mentioned, and there will be an emphasis on the need for concurrent information communication by outlining motivating scenarios. Finally, the research questions of this paper are listed.

1.1 Voice-based Interaction

In a voice-based interaction method, users interact with the system using 'voice'. According to Kortum (2008), Voice-user interaction is the script to a conversation between automated system and a user. At the system end, the components like machine listening, speech recognition, and dialog systems are involved, whereas, at the users' end, users listen to the response from the auditory display of the

system, and comprehend information using auditory perception. Figure 1.1 illustrates an abstract interaction flow in voice-based interaction. In typical voice-based interaction cycle, first, a user inputs to the system using a voice that the system understands input using state of the art speech recognition techniques, and after that in response to the input, system utters the ordered response, which is either orderly stored or instructed to be played logically to the user.

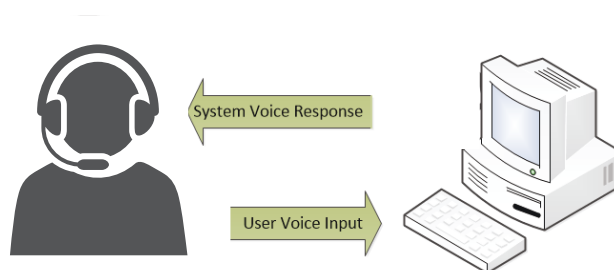


Figure 1.1 : **Voice-based Interaction:** a typical interaction flow between a user and a computer.

Our research is focused on the system response side regarding communicating multiple information concurrently in voice-based interaction. Presently, almost all of the speech-based systems such as screen-readers, text-to-speech synthesisers, audio streams, video streams, intelligent assistants, and interactive voice response (IVR) systems, communicate information sequentially to the users that create a significant issue of taking a high amount of time to reach to the point of interest by the users. Inadequate progress is made regarding transforming modern auditory display design solutions into systems interfaces (Towers, 2016). The speech *"has a low information transmission rate for continuously changing variables relative to the bandwidth of the human auditory system"* (Hayward, 1994, p. 3). This lapse is explained by evaluating an IVR-based case study in Section 1.2.

1.2 Interactive Voice Response System: PTCL Case-Study

An Interactive Voice Response (IVR) System is typically an automated telephony system that provides speech-based information to a user in an automated way (Asthana et al., 2013). In an IVR, users give input either using a keypad or via voice commands (Asthana et al., 2013). Such cost-effective systems are in widespread use, particularly, in commercial organisations (Asthana et al., 2013) to serve demanding applications, like flight reservation and telebanking, with high customer satisfaction (Kortum, 2008).

The most common IVR application in Pakistan is the PTCL (Pakistan Telecommunication Company Limited) helpline that users access to source an individual's phone number, or lodge a service complaint. For any of the purposes mentioned above, when someone dials in an IVR-based helpline number 1218, a persona in the form of prerecorded message utters a message as per script and provides different options at each stage of the process. At each stage, the user enters a key by following the spoken instructions uttered by the system to lodge the complaint or make an inquiry. A hierarchical tree-based Figure 1.2 illustrates the steps involved in submitting an inquiry in the 1218 PTCL helpline IVR system.

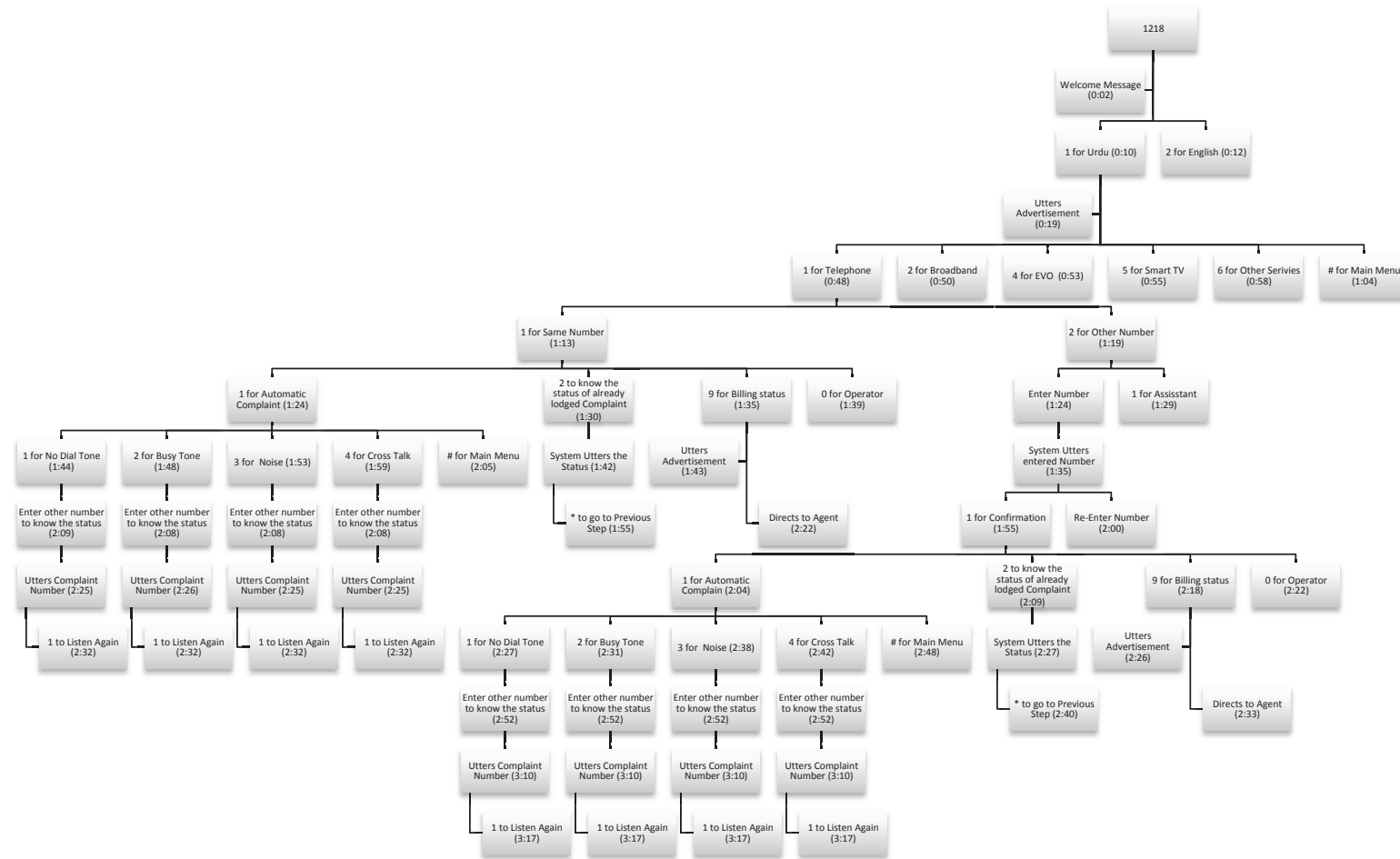


Figure 1.2 : **IVR Case Study:** Sequence of voice-based responses uttered by the PTCL helpline IVR indicating steps to a user for completing a task.

In the PTCL IVR tree, to reach a particular destination or complete the desired action, users need to move step-by-step as the system communicates information sequentially. At each step, the user is provided with options that they may select by pressing the corresponding key on the keypad. To reach another target or to find other information, users have to go back to the root and then repeat the process of following the sequential instructions.

Since this IVR system does not provide various information in parallel to the users, therefore, the users remain unable to get the gist of other relevant contextual information. Moreover, there is no method provided to the users to reach the last nodes or options directly without listening to the subsequent utterance of the system carefully. That is unless they are the frequent users of this IVR system and remember the corresponding digit(s) to reach their desired destination.

Contrary to such speech-based sequential information presentation, if the same navigation structure was presented in the visual form, it would have enabled users to glance all the options simultaneously on each step to make an informed decision to reach the desired option from the options presented simultaneously. The simultaneous presentation would have made it less time-consuming for the users to achieve the desired goal compared to a sequential form of speech-based information communication.

Besides IVR, there are many other case studies of interactive voice systems, for example digital audio streams, screen-readers, voice messaging etc. where concurrent presentation of information could be involved. To provide information concurrently in speech-based information communication, the pertinent question is, does human listening abilities support concurrent presentation?

1.3 Human Abilities of Listening to Multiple Information Simultaneously

Users are capable of listening, noticing and comprehending concurrent information simultaneously as they do in vision (Dix, 2003). They are capable of focusing their attention on the information stream of their interest when they receive competing information in parallel. The known example highlighting this phenomenon is the cocktail party problem where a person listens to multiple voice streams concurrently and manages to pay attention to a particular stream using the selection and attention abilities by prioritising the interest (Cherry and Taylor, 1954).

For auditory scene analysis, it is important to know that how the auditory system organises information into perceptual “streams” or “objects” when competing signals are delivered to the user. The auditory system groups acoustic elements into streams, where the elements in a stream are likely to have come from the same object (bre). Using this principle, sounds that have the same frequency would likely be grouped into the same perceptual stream. Hence, the selection and attention can be met and enhanced using perceptual dimensions, shown in fig 1.3, e.g., giving distinct pitch or spatial difference between the streams etc. This notion is discussed in detail in chapter 2.

Moreover, the American Speech-Language-Hearing Association has identified the central auditory process as the auditory system mechanisms and processes responsible for the following behaviours (also illustrated in Figure: 1.4):

- Sound localisation and lateralisation: the user is capable of knowing the space where the sound occurred
- Auditory discrimination: the user has the ability to distinguish one sound from another
- Auditory pattern recognition: the user is capable of judging differences and

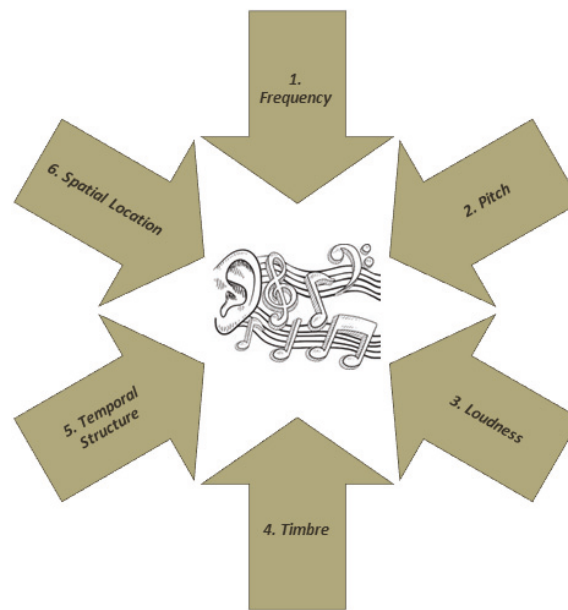


Figure 1.3 : **Auditory Perceptual Dimensions** that help in grouping the voice streams.

similarities in patterns of sounds

- Temporal aspects: the user has abilities to sequence sounds, integrate a sequence of sounds into meaningful combinations, and perceive sounds as separate when they quickly follow one another
- Auditory performance decrements: the user is capable of perceiving speech or other sounds in the presence of another signal
- Auditory performance with degraded acoustic signals: the user has the ability to perceive a signal in which some of the information is missing.

The behavioural characteristics suggest that human auditory perception has remarkable capabilities that are somehow not exploited in current voice-based human-machine interaction implementations.

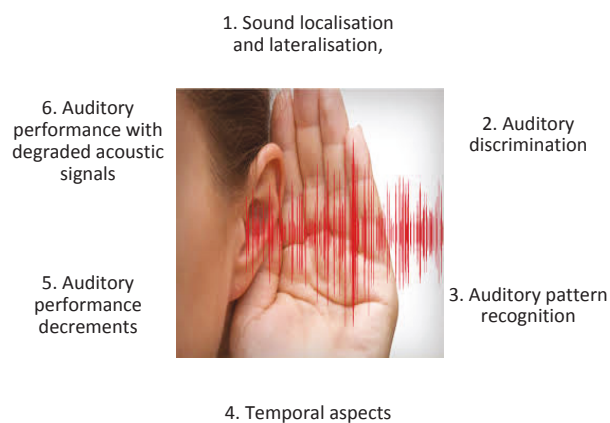


Figure 1.4 : **Auditory System Mechanisms and Processes** responsible for the different behaviours.

1.4 Multiple Information Communication in Voice-based Interaction

One of the primary goals in information design is rapid dissemination with clarity. Since user's information needs are growing (Church et al., 2014), therefore, information must be efficiently designed, produced and distributed, so that users could quickly interpret and understand. Typically, the reason for information design in any mode of interaction is to help users accomplish the goals or tasks at hand efficiently. Therefore, the human capabilities to process information should be considered (Dix, 2003) to design information (Pettersson, 2012).

In contemporary implementations of voice-based interaction, the question appears; are current information designs effective, and optimally utilising the human auditory capabilities in voice-based interaction? Unfortunately, the answer to this question is 'no', as reflected in the case study mentioned above in section 1.2. However, contrary to voice-based interaction, the visual interface provides multiple information in parallel to the user using various methods, one of which is the use of overlays (Neil, 2009). An overlay is a transparent virtual layer,

conceptually similar to the map to describe its features (Weber and Phillips, 2004). However, in reality, it could be more transparent to provide more flexibility than map transparency and simplifies the user interface providing access to additional information without leaving the page (Neil, 2009).



Figure 1.5 : **Overlay Example:** Facebook wall presenting multiple information streams on the visual interface using overlays.

Figure 1.5 illustrates the notion of overlays using a Facebook wall where multiple information is being presented concurrently. In this illustration, the overlay in the bottom right corner provides a view to the user's chat messages. The overlay at the top is showing the user's notifications. The right side pane lists the activities of friends. The left side pane displays favourites and other useful links. Moreover, on hovering the mouse over a friend's name, a preview of the friend's wall displays in another overlay. Hence, multiple information within the context is displayed using overlays where the user is able to interact with the overlay that contains the most relevant information, while other overlays are ignored.

The same concept may be adopted in voice-based interaction for communicating multiple information concurrently because the human auditory system is capable of performing filtering of the sounds received and allows users to ignore the extraneous noise and concentrate on relevant information (Dix, 2003). For

example, a subtle possibility of providing concurrent information could be to provide parallel information by broadcasting two voices concurrently, one as a primary stream representing the main information, and the other stream, as an assistant that provides additional information based on the context and behaviour Sato et al. (2011). However, the designing of such concurrent information streams can be a real challenge that would decide whether such communication method is helpful to the users, or distracts users in interacting with the system. There can be many other applications of concurrent information communication Guerreiro (2016a), some of which are discussed in the following motivating scenario section.

1.5 Motivating Scenarios

From many of the less critical applications, such as concurrent speech synthesizers, IVR, and seeking audio/video information etc., listening to two concurrent streams and gaining a gist from multiple information streams concurrently could be popular among the wider population. For example, a person might have an interest in multiple topics and have a preference for listening to live talk shows that focus on different topics. That user might be interested in listening to more than one live program at the same time, such as a talk show discussing politics while listening to a program that discusses music. This can be facilitated by concurrent information communication to save time for the user. Such information seeking could be possible in a way that a user opens two web browsers and plays both audio streams simultaneously, listening to both programs in parallel. This listening approach may be a challenging and complex task for the user who might opt to keep the volume of one stream low and the other volume high so that they can focus on the primary program and receive the gist from the secondary program. As a topic of interest is raised on the secondary program, it becomes the primary and the volume is increased and vice-versa. The higher volume is expected to help the user in keeping the focus on the primary program while the

secondary program may continuously give the user feedback or the gist.

A few other activities among the wider population that motivate one to explore concurrent communication are:

1. Students engaged in study may have multiple screens at hand. While studying, the student might have their laptop that they are working on, their phone in arms reach and a television/radio/stream playing in the background.
2. Parents who are obliged to have a child's program playing on a large television screen, while they have their programming on a smaller (less audible) screen. The parent is likely to be attempting to pay attention to both streams to ensure that appropriate content is playing on the television screen for the child while being entertained by their programming choice.
3. Video game players may have instructional video streaming on one screen while they are gaming on a second screen.

Besides the less critical applications of concurrent speech-based information communication, many other critical real-life domains may benefit from concurrent designs. Professionals who engage themselves in listening to multiple talkers simultaneously, such as air traffic controllers and physicians working in an emergency ward, who balance their responsibilities and tasks by listening and interacting with multiple sources simultaneously (Walter et al., 2017) may benefit from concurrent designs. In the medical industry, research is already heading where auditory displays enable the head-up monitoring of the patient during theatre operations (Sanderson, 2006). Similarly, possibilities of non-speech concurrent communication have also been explored regarding the airplane-deck (Towers, 2016). Concurrent speech-based information communication in such critical fields would require careful considerations and research.

1.6 Research Aim

In this research, the aim is to investigate the possibilities of communicating multiple speech-based information streams concurrently. In this investigation, the following research questions are explored.

1.6.1 Research Questions

- Is concurrent information communication possible in voice-based human-computer interaction (HCI)?
- Do users have interest in concurrent information communication?
- What type of interaction approach do users want to interact with the systems communicating concurrent information?
- How different does the concurrent information communication perform than the doubled playback-rate?
- What could be the effective design(s) for the concurrent information communication?
- How much information can be comprehended in concurrent information communication?
- What could be the optimal combination(s) of multiple information streams for concurrent information communication?

In this thesis, Chapter 2 contains the literature review, Chapter 3 discusses the first study and answers the first research question. Chapter 4 discusses the second study and addresses both the second and the third research questions. Chapter 5 discusses the third study and compares concurrent information communication with the high playback-rate in order to answer the fourth research question. Chapter 6 presents the analysis that answers question 5. In Chapter 7, information

comprehension depth by the users is determined to answer the sixth research question. Finally, Chapter 8 investigates various concurrent combinations of information types and answers the final research question.

Chapter 2

Literature Review

This chapter presents the important aspects of human auditory perception, explicitly focusing on speech perception and a range of psychological studies that have explored human abilities to comprehend concurrent information using speech. Discussing these two points assist in gaining an understanding that humans are capable of noticing, listening and comprehending multiple voice streams simultaneously, and that there is a potential of communicating multiple information concurrently. After discussing auditory perception and psychological studies, state of the art regarding concurrent information communication in human-computer interaction (HCI) is mentioned that shows how computer researchers have tried to exploit this human ability for providing quick interaction with the system. At the end of the chapter, the interest in the study and what experiments would be conducted to investigate concurrent speech-based communication is mentioned.

2.1 Human Auditory Perception

Before going into the details, it is appropriate to briefly discuss the anatomy of the sound reception in the human ear that is the primary organ of the auditory system. Besides the ear, the other main organ is the central nervous system that is responsible for the processing of the sound (Akram, 2015). As shown in Figure 2.1, the human ear consists of three parts, the outer ear, middle ear and inner ear (Møller, 2006). The visible part of the ear is called the outer/external ear that consists of the auricle and the ear canal. The middle ear has the membrane and three ossicles: malleus, incus, and stapes. The inner ear consists of the semicircular

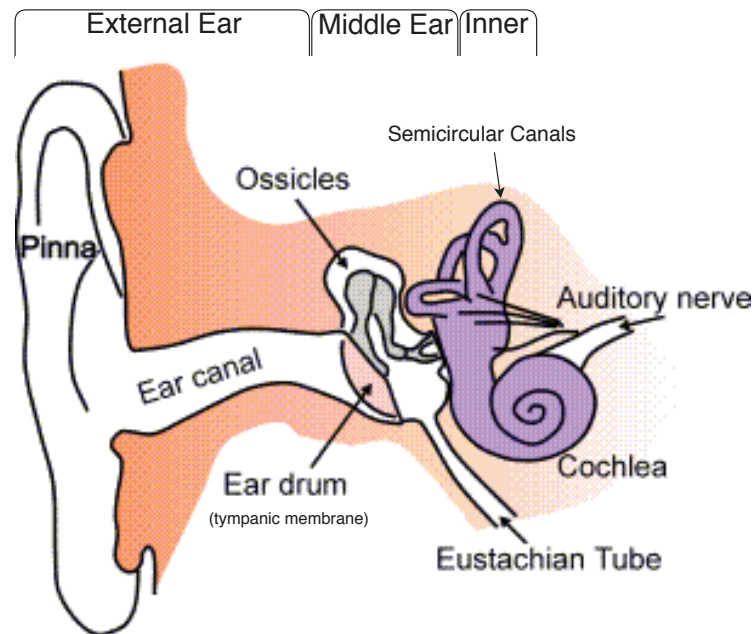


Figure 2.1 : Schematic View of the Periphery Auditory System (Clopton and Spelman, 1995; Akram, 2015) - *Public Domain*

canals of the vestibular system and the cochlea. The outer and the middle ears using the ear canal, and sometimes through bone, conduct the sound to cochlea for hearing. The cochlea separates the sounds according to their concerning frequency before it transduces into a neural code in the fibres of the auditory system. The hearing cells in the cochlea are of two types: inner hair and outer hair cells. In the inner hair, sensory transduction takes place, and the outer hair cells act as 'motors' responsible for reducing the effect of friction on the motion of basilar membrane. The separation of sounds by the cochlear activates different populations and auditory nerve fibres. The abundance of branching of auditory nerve fibres and fibres of auditory tracts are the basis for parallel processing of information in the auditory system. The information separation (stream segregation) takes place in the associated cortices where different kinds of information remain present in anatomically different populations of neurons (Møller, 2006).

In daily life, a human hears multiple sound streams from a variety of sources simultaneously. Listening to and comprehending one stream (*target*) by segregat-

ing it from competing sounds (*masker*) is known as “the cocktail party problem” (Cherry and Taylor, 1954). The segregation of streams from incredibly mixed sound or in reverberant conditions is a complex and demanding task that human auditory nervous system performs seamlessly (Cusack et al., 2004; bre). Though sound processing in a brain appears to be effortless, the underlying mechanism of sound processing at different stages in the brain could not be fully understood. Many scientific engineering approaches have been introduced to mimic the processing of sound, but they have yet remained unable in computational implementations. In previous decades there have been studies on the behavioural and neural mechanism for auditory processing and stream segregation, but the mechanism is not entirely explored (Timothy D. Griffiths and Jason D. Warre, 2004; Elhilali, 2017; Elhilali and Shamma, 2008).

According to bre the mental representation of the surroundings forms from the processing of information provided by the human senses. This mental image of the surroundings is called perception that can be divided into two processing cycles, 1) *sensory input representation* 2) *grouping of input*. In the first cycle, the informational cues and the acoustic features are extracted from the auditory scene. In the second processing cycle, the extracted features and cues are grouped to form meaningful and segregated auditory streams. The second processing cycle is extensively tangled task that the human brain performs optimally.

The brain, as a dynamic physical system, works on a set of perceptual principals. These principals are strongly influenced by the listener’s language, auditory memory, musical background, and many other factors that are common to a group of people, help the brain identify meaningful events out of limited or missing information provided in the auditory scene. In the field of psychology, the auditory scene analysis (ASA), coined by the renowned psychologist bre, is considered as a model for the foundation of the auditory perception. This model outlines a process of how the human mind organises and groups meaningful elements from sound

to make it perceivable. The determination of the perceptual boundaries between simultaneous auditory streams or to separate the auditory cues and features of one stream from the rest of the auditory scene is a challenging task. Many grouping principles have been proposed to group the relevant perceptual elements in a complex auditory scene (Bizley and Cohen, 2013; Timothy D. Griffiths and Jason D. Warre, 2004). These principles are primarily based on parsing the acoustic patterns in time-frequency space. Gestalt psychology (Köhler, 1947) is the primary source for many of the grouping principles that helps auditory scene analysis (Towers, 2016). According to Gestalt psychology, meaningful perception is acquired from the holistic view of the stimulus from the environment. Gestalt laws, such as continuity, similarity, closure and symmetry are relevant to auditory perception when viewed in a temporal form.

Sequential and parallel grouping can be viewed as two aspects. The first aspect relates to linking the spectral components to the associated sources over time whereas the second group determines linkage that which parts of the complex acoustic scene belong to which source. The principles of continuity, similarity, common motion, and proximity help in grouping the input cues (Köhler, 1947). Principles of location cues, pitch and spectrotemporal properties aid in grouping the auditory streams. Another important feature that plays a vital role in grouping the objects is harmonicity. The fusion of frequency components with harmonic relationship forms a single pitch and renders a unique entity to a harmonic complex that enables to identify the stream from a complex signal with varying fundamental frequencies F_0 (Duijhuis et al., 1982; Rasch, 1978). Moreover, in comparison to the offset synchrony, onset synchrony is considered to be more useful in grouping the cue (Darwin and Carlyon, 1995).

Sound frequency plays a fundamental role in segregating two audio streams. The two audio streams fall in the same cochlear channel in the peripheral auditory system when there is a low-frequency difference between them. In an experiment,

it was observed that two frequency tones A and B were considered to be one stream 'A-B-A-B' when the frequency separation was small (less than 10%) or low presentation rates (Rose and Moore, 2000; van Noorden, 1977). However, in the case of big frequency separation and high presentation rates both the tones perceived as two separate streams, A-A-A... and B-B-B...(Rose and Moore, 2000; van Noorden, 1977).

Another important factor in segregating the audio stream is sound localisation. In an environment, the sound sources have varying spatial locations where all the objects associated to one source share the same spatial cue. In the case of the cocktail party problem where more than one talker are speaking simultaneously, it is ascertained that the binaural listening helped significantly in stream segregation (Yost, 1994). The segregation of competing sounds in noise improves by hearing it with two ears as the central nervous system can detect 1) *interaural intensity differences*, that is the differences in loudness at the two ears and 2) *interaural time differences*, that is the time difference of sound arrival at one ear versus the arrival at the other. The ear nearest to signal is ipsilateral and the ear furthest to signal is known to be contralateral.

The other well-studied phenomenon of spatial hearing in concurrent listening is the precedence effect that is related to accurately localising the sound. Source localisation is the ability to determine the source of the audio stream. For example, when two streams reach the listener almost at the same time, and the listener perceives it as a single fused image directed from the near location of the first-arriving stream (Shinn-Cunningham, 2015). A space called soundstage by Audiophiles within which voice sources are perceived. The quality of the soundstage depends on the type of sound sources and relationship among them. The head-related transfer function (HRTF) is the spectral filtering of the sound that occurs when sound interacts with pinnae (outer ear), head, and upper torso (Towers, 2016).

The directional hearing aided by the two ears helps to select one speaker

when multiple speakers are simultaneously speaking nearby the listener. In a competing voice signal, the *target* is the voice stream that a user wants to listen to, whereas the *masker* is the background noise that may prevent a user from listening to the target. In a group discussion where many people talk to each other, the conversation holds in a highly interactive manner. People talk to each other, divert attention to the different talker, one speaker might cut another short and start speaking, sometimes more than one speaker talks at the same time that generates masking, eventually reduces the perception of the targets. In some scenarios, the environmental sound would be very high that generates a low signal to noise ratio makes the target difficult to listen to. The signal to noise ratio would be read 0dB when the background voice and the target voice reaches to the listener's ears at the same intensity. In the competing voices, all the listeners should be able to parse the different talkers. According to (Cherry and Taylor, 1954), listeners use physical differences among the competing streams to select the target. These differences include the gendered voice, intensity of the voice and the location of the voice. Hawley et al. (2004) identified four cues that are particularly relevant to the identification of the target in cocktail party problem. The cues include spatial release from masking, temporal properties of the masker, Fundamental Frequency (F_0), and informational & energetic masking.

For a stream selection from the competing streams (van Noorden, 1975) and build-up of a stream (Rosee and Moore, 2000), the role of attention has been widely studied. The studies showed that attentional focus plays a critical role in detecting the target in the presence of maskers as well the spectral separation difference between the target and maskers (Elhilali et al., 2009; Micheyl et al., 2007; Gutschalk et al., 2008). However, it could not be established concretely that focused attention plays a core role in segregating streams. In primitive segregation, the learning parameters and the attention to the sound do not play a role in segregation of streams (bre). This segregation refers to a bottom-up pre-attentive auditory process

of the auditory perceptual organisation. In the top-down or schema-based auditory processes, affected by listeners preceding experiences and acquaintance with the presented sound, active attentional state, learning, and memory play a significant role in segregating the streams. The study (Cusack et al., 2004) indicates that attention plays a significant role in the build-up process of the streaming signals. The absence of attention or divided attention reduces the abilities to report streams (Carlyon et al., 2003).

Sound signals contain an enormous amount of information. So much so that it might not be possible to process all the information, therefore selective attention plays a core role in comprehending information. The selection of a stream to pay attention to can be a random or a non-random process. In psychology, a contentious issue is when the selection of stream happens from multiple streams. Does it occur in the early stage, or is the attention decided at the later stage of receiving the sound stimuli? The foundation of this controversy traces to the dichotic listening experiments (Westerhausen and Kompus, 2018). The dichotic-listening paradigm is a widely used behavioural task for assessing hemispheric asymmetry for speech and language processing. In a dichotic-listening experiment, two different stimuli are presented where one stimulus is presented to the left, and the other is presented to the right ear of via headphones (Bryden, 1988; Westerhausen and Kompus, 2018). The participants identify and report the content they hear in each ear. In such a presentation, participants report more content from the stimulus that is presented in the right ear than the left ear. In a dichotic presentation, the initial perceptual representation can be altered by features, i.e. inter-channel onset asynchrony (time lag), inter-channel stimulus intensity differences, and trial-to-trial stimulus repetition.

In experiments conducted by Cherry and Taylor (1954), users were able to notice the change in the pitch, their names, amongst others, from the unattended stream. The early selection of attention is believed to happen after processing a lit-

the information whereas for the late selection the information is first comprehended semantically and then the selection takes place.

Models of attention predict that a human cannot purposely process all information streams in a complex acoustic scene (Jay and Gordon, 2005; Treisman, 1964; Deutsch and Deutsch, 1963). Broadbent (1958) experimented with dichotic listening to determine the functioning of attention that was going internally into one's head. This experiment introduced a filter model and concluded that humans could pay attention to only one channel (ear) at a time. Broadbent argued that since a human has limited information processing abilities, the internal filter mechanism prevents the information processing system from becoming overloaded. Treisman (1964) validated *early filtration* (Broadbent, 1958) as an important component of the auditory process. However, she argued that the 'early filtration' instead of eliminating the message, attenuates the message. Through her study, Treisman argued that the messages' processing in the brain begins with the physical characteristics analysis, syllabic pattern, and individual words followed by the grammatical structure and meaning assessment. The processing and analysis of a message require some degree of intensity reaching the threshold. The attended message and some of the attenuated items would remain successful in reaching the threshold. Some of the items from the unattended message would always have reduced threshold, for example, name and phrases like help, fire or the keyword that carries the potential information.

The Deutsch-Norman Model proposed another layer of filtration based on meaning (Deutsch and Deutsch, 1963). Deutsch and Deutsch argued that the un-shadowed message does not process in the working memory that somehow negates the Treisman Model (Treisman, 1964). Deutsch and Deutsch proposed that before reaching working memory, a message has to pass through two filters after pattern recognition. If the secondary stream is deemed unimportant, the second filter will not allow it to enter into the working memory. By this principle,

only an unattended message that has immediate important information would be processed in working memory. Some other renowned attention models are: the *multi-mode model of attention*, and *kahneman's capacity model of attention* (Jay and Gordon, 2005). All discussed models indicate the potential of communicating two information streams concurrently as humans have a remarkable selection and attention ability and as well as divided attention ability. This may help them to receive information from the computer in the same way as they get in real life, for example shifting attention towards the information source carrying high interest.

2.2 Psychological Studies Exploring Concurrency

For more than 50 years, researchers have debated whether, or not, the human brain is capable of processing at least two voice streams simultaneously. Past research studies (Yost, 1997; Arbogast and Kidd, 2000; Cherry and Taylor, 1954) indicates that the user performance is comparatively better when asked to listen to the source from one location and, the performance deteriorates when the user tries to listen from the unexpected locations. Recent studies have shown that when listeners were asked to listen to two simultaneous messages, they did remarkably well in listening to both the messages. Studies, (Rivenez et al., 2006; Conway et al., 2001; Cowan, 1998; Lawson, 1966; Moray, 1959) have reported that a listener has the capacity to process the secondary information present in messages outside of immediate focus. A listener can selectively read out the secondary information from the temporary buffers after the messages end (Best et al., 2006; Conway et al., 2001).

Iyer et al. (2013) carried out experiments to understand the amount of information storage and the nature of semantic processing in the memory for later recall. In this study, two different stories were presented to the users to determine whether the users understand the secondary information in detail, or understanding is restricted to the main idea only or the information is completely missed. For these

findings, the Iyer et al. (2013) introduced three content attention approaches: 1) the directed condition in which user listened to one story and answered the yes/no question from the same story, 2) undirected condition in which the user was asked question from one or both of the stories, 3) misdirected condition in which user attended a story, however, the questions were asked from the other unattended story. The results of these experiments identified that the users were able to grab the main idea of the unattended story that was more than getting information based on mere chance. However, it was noted that the performance for misdirected information remained significantly lower than the directed attention condition. The outcome of these experiments was found to be consistent with studies of the visual gist processing that suggests the auditory system receives the global features before diverting the attention to a particular information stream. This phenomenon provides the opportunity to introduce a Graphical User Interface (GUI) overlay concept in the voice-based response system.

Aydelott et al. (2012) used the dichotic sentence priming paradigm to determine the effect of competing messages on auditory semantic comprehension. The priming paradigm is usually used in research to explore the hemispheric differences in aural semantic processing. In this study, the target words' lexical decision performance was compared in strongly and weakly biasing semantic contexts when the words were presented in spoken sentences. The target was presented in either the left or the right ear, in isolation, and with a simultaneous competing for meaningful or unintelligible single talker of the same gender that randomly played either in the same auditory channel used by the target or the other auditory channel. The study concluded that the effect of the competing signal on the semantic processing of the words depends upon 1) the attentional requirements of the listening conditions, 2) the significance of the content of the competing signal, 3) hemispheric asymmetries in the processing of speech and semantic information (Aydelott et al., 2012). The competing signal presented to

the same auditory channel where the target onset eliminated the facilitation of congruent targets. However, the dichotic presentation of the competing and target signal improved the priming effect significantly. The meaningful competing signal at 0dB of Signal to Noise Ratio (SNR) produced priming effect only when the competing signal was presented to the right ear that remained consistent with the right ear advantage for intelligible speech.

In another experiment (Aydelott et al., 2015), the semantic priming paradigm was used to explore whether the unattended message is processed semantically or not in dichotic listening. The results indicated that the semantic processing of the unattended speech is significantly dependent on the intensity of the speech signals. The study found that the priming effect of the unattended speech only happened when the SNR was 0 dB, however, in attended target and attended primes scenario the robust processing was noticed at both the 12 dB and 0 dB SNR. The study established that the relative increase in the intensity of the competing signal would activate the semantic processing.

Comprehension or the processing of the information in the parallel sources is significantly contributed by the spatial cues (Best et al., 2006; Conway et al., 2001). A listener reports the voice by attending the primary voice using spatial cue and read-outs the secondary message from memory. Ihlefeld and Shinn-Cunningham (2008) determined the impact of the target location in two masker settings on the ability to extract information from two messages presented at the same time. In the experiment, one message was kept at a fixed level whereas, the other message varied from equal to 40 dB less than the fixed level message. The results indicated that the spatial separation of the competing messages improved the divided listening task of the user. The spatial separation improved the intelligibility of the less intense talker. This separation helps in three aspects; 1) hearing the parts of the source that would otherwise be masked, 2) grouping the signal into streams, 3) selecting the less intense talker. It was found that the more intense talker

was not aided by the spatial separation which suggested that the processing of high-intensity message process differently.

The spatial attribute of the audio stream acts like a spotlight that implies that if the streams could be kept in the spotlight, then all the streams would be capable of being processed. Otherwise, anything outside of the spotlight would be rejected by the auditory processing system. This hypothesis was explored in Best et al. (2006), in which two experiments were conducted. Consistent with the attentional spotlight hypothesis, the results of the experiments suggested that the spatial separation between the sources increased the intelligibility of individual sources in a competing pair but raises the cost concerned with having two process sources at the same time. Xia et al. (2015) identifies the impact of spatial separation between the audio streams on cognitive load and also determines that how a person's hearing impairment interacts with cognitive load concerning to multi-talker environment. In these experiments, visual tracking by a user was measured under four conditions of the multi-talkers. These conditions include 1) gender and spatial location, 2) gender only, 3) spatial location only, and 4) neither gender nor spatial location. The results showed that the spatial separation of 15 degrees between the streams reduced the cognitive load. In the case of hearing impairment, the spatial separation of the 60 degrees helped in causing lower cognition load on the listeners. The results of the experiments indicate that the measurement of the cognitive role in establishing the spatial separation cue for multiple information communication could be valuable.

The same voice, with the similar pitch and same source location, hinders the stream segregation that eventually affects the lexical analysis of the unattended streams. To see the effect of the difference in fundamental frequency F_0 range between attended and unattended messages, three experiments were conducted by Rivenez et al. (2006). The priming paradigm was involved in these experiments to detect the word related to a category presented as an attended message. The

primes were presented as unattended messages. The results showed that the detection was increased to 25 ms when there was a difference in the fundamental frequency of the F_0 ranges of both. Another study showed that two to eight semi-tones difference in fundamental frequency (F_0) provides a 5-dB benefit for buzz maskers and for masking sentences it provides a 3- and 8-dB benefits (Deroche and Culling, 2013). The nature of the masker seemingly determines intelligibility of voice that increases abruptly with small F_0 or gradually toward larger F_0 . The high frequencies also help in accurately localising the talks. Additionally, the perceived difference in frequency eventually helps in solving the cocktail party problem. Carlile and Schonstein (2006) established that the high frequencies assist in the spatial release from the masking. The experimental study also identified that the low-frequency energy also contributes to the spatial release from the masking when it is at the fundamental frequency of the talker over and above the perception of the fundamental frequency.

To determine the impact of the maskers, Iyer et al. (2010) conducted an experiment that introduced three types of maskers. In the experiment, they involved 1) contextually relevant speech based masker, 2) contextually irrelevant speech based masker and 3) the non-speech masker to examine the impact. The results showed that the multi-masker penalty appeared when the following two conditions were fulfilled. First, the soundstage has at least one contextually relevant masker that creates the confusion with the target (Informational Masking) and second, the Signal to Noise Ratio (SNR) of the target is less than 0 when the stimulus of the maskers is combined. In another experiment, Iyer et al. found that the listeners were able to detect and hear the keywords from all three talkers even in the situation where the multi-masker penalty occurs. Ihlefeld and Shinn-Cunningham (2008) explored masking impact and found that the energetic masking and informational masking's relative influences change as a function of the target to masking ratio. The results of the study validated previous researchers' findings

that the different attributes of the competing voice help to select and focus a target from the soundstage. These attributes also contribute to linking short-term segments across time. This finding encourages one to investigate the possibility of multiple audio streams simultaneously.

The number of talkers in a competing signal is another critical factor that impacts listener's performance (Freyman et al., 2004; Bronkhorst and Plomp, 1992; Pollack and Pickett, 1958; Miller, 1947; Carhart et al., 1969; Brungart et al., 2001; Yost et al., 1996). Kawashima and Sato (2015) carried out a study to determine the numerosity of the concurrent speech streams presented simultaneously. There were a total of four experiments conducted with different combinations of 1 to 13 talkers; 1 to 6 different locations; and a duration of 0.8 s, 5.0 s and 15.0 s. The results showed that the numerosity judgment depends on the ability to segregate the talkers from the speech signal. The auditory world may consist of 3 to 5 talkers at the same time depending on the listening context as it could be difficult for the listener to distinguish more than three to five streams reliably from the concurrent speech. The spatial difference between the streams significantly improves the numerosity judgment of the concurrent voices. In the monitoring task when the number of speakers increases, unsurprisingly the performance decreases. Researchers in many studies have identified that when the number of maskers goes beyond two, then the impact of the newly added masker does not decrease the performance significantly. Simpson and Cooke (2005) found that when single masker was involved in the signal, the accuracy was 85%. However, when the masker increased to two, the performance reduction was noted 17% and when the third masker was added only 8% reduction was noted additionally. Another study found that the amount of masking was 8 dB when the second interferer was added and when the third and fourth interferers were added, the reduction in performance was noted to be 3 dB only (Miller, 1947).

The working memory capacity (WMC) plays a significant role in focus and

attention abilities of a user which is crucial for concurrent speech processing. A listener having higher WMC keeps better control in focus and attention compared to the one who has low WMC (Yu et al., 2014). To establish this, Yu et al. (2014) used the dichotic listening paradigm where users were asked to attend the words presented to one auditory channel and ignore the other speech signal presented on another channel. In Kane et al. (2001), 65% of listeners who had a low WMC reported their name when it was spoken in the unshadowed auditory channel, whereas the 20% of the listeners who had higher WMC reported their name. It reflected that the listener with higher WMC has greater control over attention diversion.

Ageing may reduce listener's comprehension abilities in competing voices, particularly for older adults (Arlinger et al., 2009; Schneider et al., 2010; Humes and Dubno, 2010). For older adults, it becomes challenging because the focus and attention abilities of a human decrease as they age. James et al. (2014) carried out a study using a dichotic priming paradigm to examine whether the differences in cognitive function of human predicts older adults' ability to access sentence-level meanings in competing speech, or not. The study showed that older people were vulnerable to interference when the competing voice presented to the right ear. The study also validated that cognitive factors play a key role in competing speech (James et al., 2014). Getzmann et al. (2016) carried out experiments where both young and aged people were asked to attend the two types of information conveying methods. In one experiment, they were asked to attend the speech from a single target speaker, and in another experiment from 2 different target speakers (divided listening). In divided attention, it was observed that the perception abilities decreased for older people. The study showed that younger listeners have productive preparatory activity and allocation of attentional resources.

Some researchers, (Aydelott et al., 2012; Westerhausen and Hugdahl, 2008; Hugdahl, 2016), showed that the right ear may provide some advantages in competing

voice streams because in dichotic listening, the signal that reaches to right ear gets direct access to the left posterior temporal lobe for speech processing in the brain, whereas the signal presented to left ear enters to the wrong hemisphere that then gets transferred across the corpus callosum to be processed (Westerhausen and Hugdahl, 2008; Pollmann et al., 2002). The dichotic listening and the REA is one of the most frequently used methodologies used in studies regarding competing speech (Hugdahl, 2016). The right ear advantage (REA) offers better reporting of the voice stream presented to the right ear compared to the other competing voice presented in the left ear of the listeners in dichotic listening (Hugdahl, 2016).

Taking technology and the quality of the audio file into account, (Lindborg and Kwan, 2015) determined that the quality of the audio stream plays a significant role in comprehending information in competing audio. Experiments were arranged to investigate the impact of audio quality in determining the source localisation. The study concluded that the interplay between the audio file compression rate and target position lead significant impact. The compression rate impact remains different on the localisation of the wide target position and the narrow target position.

The above discussed psychological studies provide a number of cues that include: spatial difference, pitch, speed, gender, audio quality, REA, type of information that can be explored to concurrently communicate multiple information through a computer-based auditory display.

2.3 Contemporary Research Studies on Concurrent Speech Interface

In voice-based human-computer interaction, auditory display is the use of sound by the system to communicate information or the state of the computer to the user (Kramer, 1994; Hinde, 2016). The auditory displays can use non-speech or speech-based messages to convey information (Hermann, 2008). In a non-speech-

based auditory display, different techniques can be used to either enhance the visual display or communicate information using audio. The non-speech-based auditory displays include audification, sonification, auditory icon, earcons, musicons, spearcons, and spindex (Tilman et al., 2008; McGookin and Brewster, 2003). Though these design techniques slightly differ with each other in representing data, the purpose of using all of these non-speech sounds in auditory displays is to convey information to the users with sound. In legacy, non-speech sonification, multiple technical approaches (Brazil et al., 2009; Brazil and Fernström, 2006; Hus-sain et al., 2015b; Schuett et al., 2014) are employed that are also considered in this work (i.e. spatial positioning, fundamental frequency separation etc). This may further allow the comparison of the results obtained with corresponding outcomes that are reported in the literature for traditional non-speech sonification.

Besides the non-speech audio, the use of speech-based messages in auditory displays seems useful, as humans in their daily life interact with each other using the same method which provides enormous flexibility and precision to exchange information. In turn, this makes speech an ideal method to be used in auditory displays for communicating information to the user (Hinde, 2016). Conventionally, the speech interfaces communicate speech-based information in a single speech stream that, as discussed above, under-utilises human auditory capabilities. A few researchers have worked on introducing concurrent communication through speech display many of which are mentioned below.

AudioStreamer by Schmandt and Mullins (1995) is one of the first auditory displays that endeavoured to use people's ability to attend the desired stream from the competing streams selectively. In this system, three concurrent speech-based streams are presented by applying spatial difference leveraging on the cocktail party problem. The streams are binaurally spatialised to 0 and 60 degrees which were considered to be enough for perceptual segregation and quick attention switching between the sources. This configuration is based on the findings of

(Rhodes, 1987), that it takes an increased reaction time for enlarged angular separation in non-speech localisation tasks. Besides the spatial separation, the streams are presented with different talkers' voices to employ acoustic variations for reducing informational and energetic masking between concurrent talkers. The system is designed to track the head movement to identify the user's interest in a stream of the competing streams. If the head tilts towards the particular source, there is a temporarily increased gain, and then the gain is steadily normalised. Moreover, for isolating a stream and making the other stream silent, a user is required to look at the virtual source twice. To make sure that the important information is not missed, the system was also configured to momentarily draw the user's attention toward key points of other streams. In his Master's dissertation, Mullins (1996) stated that *AudioStreamer* users were overwhelmed by three channels of concurrent speech. To overcome this, Mullins introduced five-second onset asynchronies between the streams. Unfortunately, there was no formal study to test this configuration. Therefore, it was difficult to say whether this intervention improved the users' experience and information communication.

Schmandt (1998) introduced *Audio Hallway* as his second auditory display exploiting the concurrent speech-based presentation that allowed the browsing of vast compilations of audio files. The system incorporated two types of navigation on the high level and the low level. In high level, users were able to navigate within the groups of clustered content, whereas, in the low level, navigation for individual files within a cluster was facilitated. Regarding the methodology adopted for the high-level navigation (Schmandt, 1998), it was reported that linking multiple spatially separated streams with listener position movement appeared inappropriate for auditory displays. Regarding the approach adopted for low-level navigation, Schmandt (1998) reported that since the auditory items were easily associable with the orientation of the listener's head, therefore, the navigation appeared less challenging for the users.

To enable auditory display of GUI programs, Parente (2008) explored a new approach where display described concurrent application tasks using a small set of simultaneous speech and sound streams. Parente (2008) carried out a study to perceive problems faced and techniques adopted by users to interact with an ideal auditory display. For this purpose, Parente developed an auditory display prototype, called *Clique*. In this system, users, instead of interacting with the underlying graphical interfaces, listened to and interacted solely with the display. Mapping GUI components supported such level of adaption to task definitions. The evaluation showed that efficiency, satisfaction, and understanding was improved with little development effort. *Clique* yielded many benefits, particularly for visually impaired users and mobile sighted users through fast and accurate access to speech utterances, better awareness of peripheral information, increased information bandwidth, effective information seeking, and faster task completion, to name a few.

The availability of digital media has transformed the means by which people find and interact with information. Visually impaired persons mostly rely on their auditory system to receive information. Guerreiro and Goncalves (2016) carried out doctoral research on blind and sighted users, and conducted experiments to determine the information scanning abilities of the sighted and the visually impaired person from the concurrent speech. Guerreiro and Goncalves leveraged the concept of cocktail party problem. Their study was conducted on 23 sighted and 23 visually impaired users. Guerreiro and Goncalves (2016) aimed to catch people's ability to *scan* important content by listening to two, three, or four speech channels played concurrently. The sound sources were separated by the angles of 180, 90, and 60 degrees for two, three, and four talkers, respectively.

As shown by other researchers discussed above in section 2.2, Guerreiro and Goncalves (2016) found that the spatial difference in sources is the best cue in concurrent speech. The study established that sighted and the visually impaired

users have the similar abilities to scan the information from the concurrent speech (Guerreiro and Goncalves, 2016). Two concurrent information streams appeared to be more useful in understanding and identifying the content. The study showed that the use of three speech sources depends on the task intelligibility demands and listener capabilities. In another study by Guerreiro (2013), it was found that the concurrent speech with slightly higher playback-rate enables a significantly quicker scanning for relevant content. Guerreiro (2013) found that gender difference in voices does not play a role in the higher understanding of the content.

Ikei et al. (2006) introduced the vCocktail design i.e., a novel voice menu presentation method for efficient human-computer interaction in wearable computing. The design introduced spatiotemporal multiplexed voices with enhanced separation cues aimed to shorten the length of the serial presentation of voice menus. In the experiment, the appropriate directions were measured by calculating the perception error in judging voice direction, and then by following spatiotemporally multiplexed conditions with several different settings of spatial localisation, the number of words, and onset interval, the voice menu items were presented. The results showed that the subjects, aided with the localisation cues and appropriate onset intervals, were able to hear menu items accurately. Moreover, the proposed attenuating menu voice and cross-type spatial sequence of presentation that effectively improved distinction between menu items further increased the ratio of correct answers.

Similarly, regarding the concurrent menu, Werner et al. (2015) compared the simultaneous aural presentation of up to seven menu items with a conventional serial aural presentation of menu items. In this simultaneous form, users were enabled to scan the auditory display to find the most appropriate command. Thirteen users participated in this study to investigate the viability of this simultaneous approach. The system, called *VoiceScapes*, appeared more difficult and attentionally more demanding compared to the other forms of presentation. However, it is

expected that VoiceScapes might allow experienced users from extended use to navigate complex menu hierarchies efficiently (Werner et al., 2015).

For temporal navigation of audio data, Minoru Kobayashi and Chris Schmandt (1997), by taking advantage of human abilities of simultaneous listening and memory of spatial location, presented a spatial interface based on a browsing environment. In this system, the user, instead of forwarding or rewinding the audio, browsed the audio data by switching attention between multiple moving sound sources played from one audio recording. The movement of sound sources mapped temporal position within the audio recording onto spatial location and users, with the help of memory of spatial location, determined the specific topic from the recording. Using the system, it is showed that the spatial memory of audio events is usable for audio browsing, and considered it a new dimension of the spatialisation technologies regarding the temporal navigation of audio (Minoru Kobayashi and Chris Schmandt, 1997).

Frauenberger and Stockman (2006) on tackling the lack of re-usable design, discussed the design patterns regarding auditory display by employing 3D virtual audio environments with concurrent audio streams and tested against the latest screen reader. They used the idea of a virtual horizontal dial with items located around its perimeter to propose a navigation design for auditory menus using concurrent speech. The results showed that the shortcomings pointed in the previous prototype discussed in the same study (Frauenberger and Stockman, 2006) were removed, but despite the improved naturalness, the marginally better performance could be achieved with the screen reader, that shows the auditory design is still an ad-hoc solution. For providing guidelines to mobile application designers to build eyes-free auditory interfaces, Vazquez Alvarez and Brewster (2010) used a divided-attention task and conducted an experiment where an audio menu and continuous podcast competed for attention. In the experiment, the impact of the cognitive load was assessed using the NASA-TLX subjective

cognitive load assessment tool. The results showed that users' ability to attend two concurrent streams enhances by spatial audio, and also the divided attention impacts the overall performance significantly.

Hinde (2016) explored how auditory displays can offer an alternative method for television experiences that depend on users' desire of being able to attend to screen-based information visually. In his doctoral theses he carried out studies to design auditory displays involving varying levels of concurrency for two use cases regarding television: 1) menu navigation, and 2) displaying relevant content besides a TV show. Regarding the navigation of auditory menus, the first study investigated spoken menus regarding word length and onset asynchrony. The study devised optimum asynchrony and showed that the better performance could be achieved with the shorter words. The second study investigated the impact of providing additional content accompanying a television program concerning disruption, workload, and preference. The results showed that offering sound-based secondary content from a smartphone after removing the speech from the television program was the best auditory approach.

For improving a pilot's situational awareness for the changing state of systems information, Towers (2016) supported the use of spatial auditory displays within flight decks. For improving a pilot's situational awareness for the changing state of systems information Towers (2016) supported the use of spatial auditory displays within flight decks and conducted four studies to evaluate an auditory display involving spatially positioned sonifications to communicate information from multiple navigation displays. The results of the studies supported the use of concurrent spatial sonifications as it helped users to spend more head-up time to an out of flight deck visual search task and fly the aircraft more precisely. Moreover, for verbal navigation instructions, the left ear, or along the midsagittal plane appeared most effective that showed left ear advantage significantly in the context of competing for attention with sonified spatial navigation data (Towers,

2016).

2.4 Summary of the Angles Reviewed

In this chapter, various aspects of human auditory perception have been reviewed, explicitly focusing on speech perception and discussing a broad range of psychological studies that have explored human abilities to comprehend concurrent information using speech. This review helped in understanding that humans are capable of noticing, listening and comprehending multiple voice streams simultaneously and that there is potential for communicating multiple information concurrently. The psychological studies provided some cues that include spatial difference, pitch, speed, gender, audio quality, REA, and the type of information that can be used to explore multiple information communication in a computer-based speech display. This chapter further reviewed almost all the research endeavours for communicating multiple speech-based information concurrently in voice-based interaction from various perspectives.

2.5 Our Research

The reviewed studies explored different angles but did not carefully take users' interest and their expectations from such systems into account. For example

- a) Would they be able to comprehend information from both the streams?
- b) Would users prefer such a system if provided with multiple information concurrently?
- c) Would users always like concurrent information over sequential presentation, or there would be some limitations and contextual needs that may involve?
- d) What are the users' expectations from such a system, and how would they want to use it?
- e) What interaction possibilities can be provided?
- f) How different the users' comprehension would remain in high-playback rate and concurrent communication compared to the baseline condition?
- g) What could be the optimum

concurrent design(s) that could help users to achieve the comprehension level that they attain in baseline sequential communication? h) When provided with the concurrent information, how much information would they be able to comprehend? i) What would the pattern of comprehension, would the pattern be similar to the baseline condition, or would remain different? j) Last but not least, what types of information streams can be best suited for concurrent communication to the users?

This thesis extends speech-based concurrent communication research and addresses the points mentioned above. For this, two pilot studies were conducted and then carried out two standardised experiments. The pilot studies were conducted to gain an understanding of the users' interest in concurrent communication and to examine whether or not users can comprehend concurrent information. In the standardised experiments, different designs for speech-based multiple information communication were tested to determine the depth of comprehension by users in each design. In the final experiment, various combinations of information streams were tested to investigate the cognitive workload. Each study and experiment are individually discussed in the following chapters.

Chapter 3

Viability of Concurrent Information Communication

In this chapter, a pilot study has been discussed that explored the viability of concurrent communication of multiple information by a computer system to the users. This study investigated whether users are capable of noticing, focusing and comprehending multiple information streams presented concurrently. A prototype design was created to investigate these questions. In prototype, an audio bulletin was built where two information streams, one continuous and the other in intermittent form, were communicated to the users concurrently. This study contributed us in designing careful subsequent studies to explore the various designs for communicating multiple information concurrently to the users so that they could fulfil their growing information needs and ultimately complete multiple tasks in hand efficiently.

3.1 Aims & Motivation

Aims: The aim of this preliminary investigation is to determine the viability of concurrent communication of multiple information by a computer system. The investigation set out to satisfy the following questions: a) What perception cues can be helpful in processing concurrent information, are they different than the cues reported in the literature? b) What would be users' behaviour when they miss some information from concurrent streams? c) How do users employ their selection and attention abilities? d) Would they prefer concurrent information communication over the sequential form of communication? e) Is the user context important for concurrent information communication?

Motivation: The motivation is that results would support us to design careful subsequent studies in order to explore various designs and approaches in the right direction for communicating multiple information concurrently.

3.2 Methodology

3.2.1 Stimulus Material

In this study, an audio bulletin was built where two different voice-based streams, one in the female voice, and the second in the male voice, were played concurrently. To build this audio bulletin, two different video bulletins of BBC Urdu's program *Sairbeen* were selected. *Sairbeen* is a renowned news bulletin that contains worldwide reports, expert opinions, public opinions, features on interesting topics and current affairs.

The video bulletins were converted into two audio files of *.wav* format. Each audio file consisted of three different news stories. From the first audio file in the female voice, detailed news about an exhibition was extracted. From the second audio file in the male voice, just the headlines of all three news were extracted and were broken into three audio files. Hence, there were four audio files, one in the female voice playing the documentary named the primary voice, and three in the male voice, playing news headlines, named the secondary voice.

3.2.2 Design

A concurrent information design strategy was used when playing the streams to users. In this design, the primary stream was continuously played in the left ear, the secondary voice was intermittently played in the right ear with a silent interval of 10 seconds between each headline. The delivery of playing information streams to different ears (dichotic listening — panning — spatial difference) was adopted to make it easier for the users to segregate both voice streams. The gendered

voices and spatial difference were involved in the stimulus design as in literature review discussed in Chapter 2 they were mentioned as two important cues for segregating concurrent streams.

All the information streams with the applied design were merged into one stereo audio file, (*listen here*, <http://bit.ly/2RQaaTs>), by writing a program in Visual Studio 2013 C#. The duration of this audio clip was 1 minute and 28 seconds. The rendered clip was played on Dell Vostro 5560 with Core i5 processor and 4GB RAM, and KHM MX earphones were used to listen to the clip. The earphones were used as they were expected render better spatial difference experience for the users than the built-in computer speakers.

3.2.3 Participants

The study was conducted with 10 users, 6 male and 4 female, with the age range of 20 to 55 years, see Table 3.1. Users participated from their workplace. The setup was taken to their places for their participation. After receiving participation consent from the users, they were briefed that the study can be conducted at random time during their routine activities. Therefore, it was not considered whether they are free to participate or they are in the middle of an alternate task. This approach was adopted to cover the ecological setting where a user could be under a workload pressure or with a relaxed mind. This approach was adopted to cover the ecological setting where a user could be under a workload pressure or with a relaxed mind. The overall results discussed below inclusively represents the mixture of both states. The audio file was played only once to each user.

Table 3.1 : **Participants Demography:** Composition of participants w.r.t. their *gender, age, and hearing disability*

<i>Gender</i>	<i>Total</i>	<i>(20-30)</i>	<i>(30-40)</i>	<i>(40-55)</i>	<i>Hearing Disability</i>
Female	4	2	1	1	No
Male	6	1	2	3	No

3.2.4 Questionnaire

In order to investigate the behaviour of users, a questionnaire, shown in Table 3.2, was prepared. The interviewees were briefed about the audio playing mechanism. Before they started to listen to the audio clip, they were provided with an overview of the questionnaire through an example or two indicating that what types of questions would be asked so that they could focus accordingly. The questionnaire was expected to establish whether a listener could notice, focus and comprehend multiple information streams simultaneously or not. The questionnaire also aimed to scale the notice, selection and attention behaviour of the user by asking questions from the played content.

Participants were given cues and three multiple choice options to answer each question in order to reduce the user's memory load. Besides the questions that related to content, the users were also asked whether they were able to listen to both the sounds and discriminate each of the voice streams. Finally, they were asked, would they prefer such information presentation over the sequential form of communication.

3.3 Results

Most users were able to answer the questions correctly that related to whether they could hear both streams concurrently, and whether they understood the content. Also, in the perceptual and observational questions, all of the participants found voice streams audible and discriminable in concurrent form of presentation.

In following sub-sections, the user responses for each question asked in the questionnaire is discussed individually. The users' answers to the questions are also briefly mentioned in Table 3.2.

1. Could you hear the primary voice presenting documentary? From all participants, 80% of the users told that the primary voice was more clear to hear. The

Table 3.2 : **Questionnaire & Users Responses:** Reflects the responses received from users against each question asked in questionnaire. Desired/correct answer against each question is shown in the first response column.

Question	Response: 1	Response: 2	Response: 3
1. Could you hear the primary voice presenting documentary?	Clear: 80%	Could be Improved: 20%	No: 0%
2. What was the topic of primary voice?	Exhibition: 100%	ISIS Attack: 0 %	Metro Bus Service: 0 %
3. Where the exhibition was scheduled to held?	Karachi: 40 %	Lahore: 10 %	Don't Know: 50 %
4. What was the venue name?	Mohatta Palace: 100 %	Sareena Hotel: 0 %	Convention Center: 0 %
5. Could you please tell us more about the exhibition documentary?	-	-	-
6. Could you notice the secondary voice?	Yes: 100 %	No: 0 %	-
7. Were you able to distinguish secondary voice in presence of primary voice?	Clear: 70 %	Could be Improved: 30 %	No: 0 %
8. What was secondary voice indicating?	News: 100 %	Songs: 0 %	Commercial Ads: 0 %
9. How many times secondary voice played in different intervals?	Three Times: 70 %	Five: 30 %	One: 0 %
10. In the first occurrence what was the topic of secondary voice?	Budget: 90 %	Weather Forecast: 10 %	Bollywood: 0 %
11. In the second occurrence what was the topic of secondary voice?	Cyber Attack: 90 %	Exact Degree Scam: 10 %	Cricket Team Tour: 0 %
12. In the third occurrence what was the topic of secondary voice?	Turkey Election: 90 %	Terrorists Killed: 10 %	Don't Know: 0 %
13. Which was the most interested news for you?	Cyber Attack: 70 %	Budget: 20 %	Exhibition Documentary: 10 %
14. Did you want to promptly listen to the detail news from any of the spoken news?	Yes: 100 %	No: 0 %	Not Decided: 0 %
15. Would you prefer multiple sounds over sequential flow of information?	Yes: 90 %	No:10 %	Not Decided: 0 %

remaining 20% of users, though said that they were able to listen to the primary voice, remarked that it was loud and shrilling. There is room for improvement to make it more understandable.

2. What was the topic of primary voice? All the participants correctly identified the topic of primary voice i.e. Exhibition.

3. Where is the exhibition scheduled to be held? Many of the users were not able to answer this question correctly. The probable reason for this users' behaviour is discussed in section 3.4.

4. What was the venue name? All the participants answered the venue name of the exhibition correctly, i.e. *Mohatta Palace*.

5. Could you please tell us more about the exhibition documentary? To investigate users' comprehension, they were asked to describe the content they listened in the exhibition documentary. All the users were able to describe the documentary and gave the overview in broken words often using the keywords from the documentary.

6. Could you notice the secondary voice? All users answered 'Yes' to this question, as they were able to notice the secondary voice in the presence of the primary voice.

7. Were you able to distinguish secondary voice in presence of primary voice and vice versa? As shown in Figure 3.1, 70% of the users stated they did not find it difficult to distinguish the secondary voice from the primary voice and vice versa. 30% of the users argued that segregation should have further facilitated in the presentation as they missed some of the information while focusing on a particular voice.

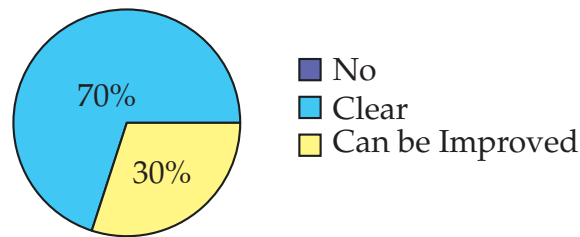


Figure 3.1 : **Distinguish Secondary Voice:** proportion of users answering this question.

8. What was the secondary voice indicating? All users correctly answered that the secondary voice was presenting news.

9. How many times secondary voice played in different intervals? 30% of the users answered incorrectly, whereas, 70% of the users answered correctly, i.e., *three times*.

10. In the first occurrence, what was the topic of secondary voice? Among all participants, only one user was not able to correctly answer this question.

11. In the second occurrence, what was the topic of secondary voice? 90% of the participants correctly answered that the topic of the second occurrence was *cyber attack*.

12. In the third occurrence, what was the topic of secondary voice? The same result was noted as was observed in the above two questions.

13. Which was the most interesting news for you? In reply to a question where users were asked to mention the most interesting news that they found among all four streams, as shown in Figure 3.2, 70% of the users opted *Data theft in Cyber Attack*, 20% of the users opted for *Black Money in Budget*, and only one user chose *Exhibition Documentary*.

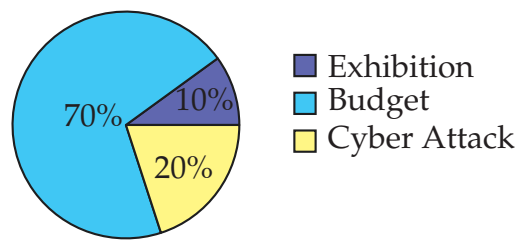


Figure 3.2 : **Interesting News Topic:** proportion of users answering this question.

14. Did you want to promptly listen to the details of the news that you found interesting? As a successor to the question 13, when users were asked to mention whether they wanted to listen to the details of the news, they found interesting, by stopping the competing primary voice, 100% of the users answered *Yes*.

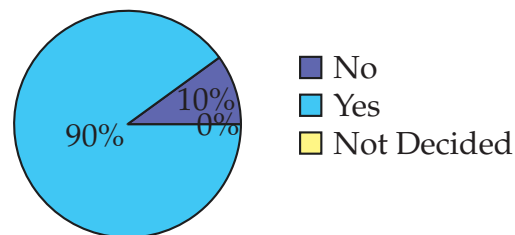


Figure 3.3 : **Multiple Information Preference:** proportion of users answering this question.

15. Did you find multiple sounds helpful in reaching multiple information quickly and would you prefer this approach over sequential flow of information? Last but not least, users were asked to answer the pertinent questions, whether they found multiple sounds helpful in reaching multiple information quickly, and would they prefer this approach over the sequential form of information presentation. User's response is reflected in Figure 3.3.

3.4 Discussion

All the users were able to describe the primary voice stream, i.e., documentary and gave the overview in broken words from the documentary. These broken words can be attributed as the keywords that contained the main idea of the information stream. This shows the importance of using keywords in concurrent information communication. Keywords are likely to have helped users to gain the gist of information from concurrent streams. Users were also able to answer the questions related to secondary voice. This user's behaviour supports the assumption that users were able to receive and process information from the concurrent streams.

Reporting content from both the streams might have been easy because users were able to segregate the voice streams. It is likely that the ease of discriminating both the voice streams was because each stream was entering different ears separately, i.e., *spatial difference*, and also the voice streams were *gendered*. Based on this, it is likely that other auditory dimensions might also benefit the process of *segregation*.

For question three (*Where is the exhibition scheduled to be held?*), many users were unable to answer this question. On investigation, a potential reason found that as the primary voice uttering the city name, the secondary voice was also being played at a higher volume than the primary voice. Therefore, this is assumed to be a reason that users could not answer the city name correctly. Those who answered this question correctly, guessed by applying prior knowledge. This indicates that whenever the users missed some of the information, they filled it from prior knowledge.

Answering question fourteen (*Did you want to promptly listen to the details of the news that you found interesting?*), users mentioned that they were keen to immediately listen to the details of news topic of interest to them. This supports that when users are provided with concurrent information they initially process

both concurrent streams and then based on the interest may divert the focus towards the interesting information stream using the selection and attention abilities. This phenomena thus provides an opportunity to introduce the same strategy that can be seen in GUI using overlay or lightbox discussed in Section 1.4.

Responding to the question asking about liking concurrent information communication, 90% of the users found this quick design of delivering information helpful and said they would prefer concurrent information communication over the sequential flow of information. From that 90% of the users, a few had reservations and mentioned that in such technique they might lose valuable information and that they would prefer to listen without any noise and disturbance (masking). They further remarked that though it might be quick to reach information, they might miss some crucial information like passwords and security codes etc. This shows that user context is important, and therefore, it could be an interesting study to find the users contexts and the types of information streams in which the concurrent information communication design could be applied.

3.5 Limitations & Future Work

The primary goal of this pilot study was to determine the viability of concurrent communication of multiple information by an auditory interface. The methodology for this investigation was not standardised. Results of this investigation are encouraging to explore the concurrent streams design approach further. The subsequent studies attempted to find the answers to the remaining research questions mentioned in Section 1.6.1. For the next study, in addition to the sighted users, it was considered suitable to investigate this information design with visually impaired users who could potentially be the greatest beneficiary of such concurrent communication method and providers of valuable feedback.

An earlier version of the research discussed in this chapter has been published in the following paper:

Publication 1: M. A. u. Fazal and M. Shuaib Karim, “Multiple information communication in voice-based interaction,” in *Advances in Intelligent Systems and Computing*. Springer, pp. 101–111.

Attached as Appendix-G.

Chapter 4

Viable Interaction Approach to Interact with the System Communicating Concurrent Information?

This chapter reports investigations conducted with 10 visually challenged users (VCUs) and 8 sighted users (SUs) that aimed to determine user's interest and expectations from concurrent information communication systems. For investigating this, two sub-studies – 2-A, and 2-B – were conducted where the participants were provided with a prototype. In the study 2-A, the prototype played two continuous voice-based information streams diotically differing by gender and content. In the study 2-B, the prototype communicated one continuous information stream in the female voice and three intermittent headlines in a male voice dichotically. This chapter then reports on participants' experience qualitatively and also perform quantitative analysis to determine users' comprehension in both the studies. Based on the experiences and feedback received from users, a framework has been proposed that may help in developing systems involving multiple voice-based information communication to the users. It is expected that the application of this new framework to information systems that provide multiple concurrent communication will provide a better user experience for users subject to their contextual and perceptual needs and limitations. To conclude, a fully functional prototype system was also developed that exemplifies the proposed framework that enables users to interact with the communication of multiple sources of information.

4.1 Aims & Motivation

Aims: This investigation aims to determine whether users, including VCUs, are interested by concurrent information communication; by what type of interaction approach users wish to interact with such systems. In addition, the investigation set out to satisfy the following questions: a) Would users always prefer concurrent information presentation over sequential presentation, or would there be some limitations and contextual needs that may mediate this choice? b) What are the users' expectations from such a system, and how would they want to use it? c) What interaction possibilities can be provided? d) Does a user's profile impact performance in comprehending multiple information? e) Do VCUs and sighted users prefer this approach and comprehend information equally or is there a difference between the two groups? f) What form of concurrent communication, i.e., continuous or intermittent, did users find more helpful in comprehending concurrent information?

Motivation: The motivation for this investigation is to understand user expectations of systems of this nature, and based on their interaction requirements to design a framework that could help designers in building interactive systems capable of communicating concurrent information. It is also expected that results would further support us to design careful subsequent studies.

4.2 Investigation

We investigated two approaches for presenting multiple information concurrently – the *Continuous* method and the *Intermittent* method, and the user's comprehension of presented information was explored in both of these approaches. This investigation further explored whether the user's educational level (tertiary or non-tertiary), played a role in comprehending multiple news-based information concurrently?

4.2.1 Participants

Ten VCU and eight SUs with a median age of 28 years, participated in these two studies. Their profile characteristics are shown in Table 4.1. All participants were well-versed in the Urdu language used in the experiment, as this language is their National language.

Table 4.1 : **Users' Profiles:** Each row indicates a user's visual *type* (Visually Challenged User (VCU) or Sighted User (SU)), whether they hold a *tertiary* qualification, if they expressed an *interest* in news & technology and the *listening* score result.

User.	Type	Tertiary	Interest	Listening
1.	VCU	Yes	Yes	10
2.	VCU	Yes	Yes	10
3.	VCU	Yes	Yes	10
4.	SU	Yes	Yes	10
5.	SU	Yes	Yes	10
6.	SU	Yes	Yes	10
7.	SU	Yes	Yes	10
8.	SU	Yes	Yes	10
9.	VCU	Yes	No	9
10.	SU	Yes	No	10
11.	VCU	No	Yes	9
12.	VCU	No	Yes	10
13.	VCU	No	No	10
14.	VCU	No	No	10
15.	VCU	No	No	10
16.	VCU	No	No	10
17.	SU	No	No	10
18.	SU	No	No	9

For recruitment of VCU participants, the National Training Center for Special Persons (NTCSP), Islamabad, responsible to train the special pupils including the VCUs, was officially contacted, letter attached as Appendix-A, and briefed about the goals of the investigation. The NTCSP obtained the consent of the VCUs, which included staff and students. Having users with varying academic backgrounds and profiles in studies helped to analyse users' performance and comprehension from multiple angles. For analysis, two user groups were organized. The grouping of users was arranged on the basis of users' qualifications specified in Table 4.2.

Table 4.2 : **Profile-based User Groups:** Two user groups based on user's tertiary qualification.

Group #	Academics	Interest	No.
G1	\geq Tertiary	Any	10
G2	$<$ Tertiary	Any	8

Participant selection involved testing normal listening abilities for both the VCUs and SUs. This test was used to ensure that the users, particularly the VCUs, are not having a hearing impairment. For this, a subtle hearing test was carried out in which three sine tones of 440 Hertz were played to the user. After one second, the tone either increased or decreased in volume to 3dB or remained flat Web-AudioCheck (2016) to check whether the users were able to notice the 3dB level difference in the volume or not. Ten sine tones were played one by one to the user and the score of correct identifications out of ten was recorded. The score for each user in their hearing test is mentioned in Table 4.1. There was no score less than 9 for any user.

4.2.2 Study 2-A - Continuous: Stimulus & Questionnaire

For this study, the prototype played the audio stream of two television shows concurrently to the user. Both audio streams were set to play continuously and diotically to the users in both ears. One stream played in the female voice, and the other stream was in the male voice. Both streams were obtained from Pakistan's largest media Group 'Geo News (Geo-News).' The topic of the article spoken in the female voice was Women Empowerment, (*listen here, <http://bit.ly/2SHbk3T>*), whereas the topic of the article spoken in male voice was on the China-Pakistan Economic Corridor (CPEC) Development, (*listen here, <http://bit.ly/2L9J7jF>*). Both news articles were in the Urdu language. The streams were played concurrently for one minute to the user who was asked to listen to both talk shows concurrently via earphones. The stimulus design for Study 2-A is illustrated in Figure 4.1.

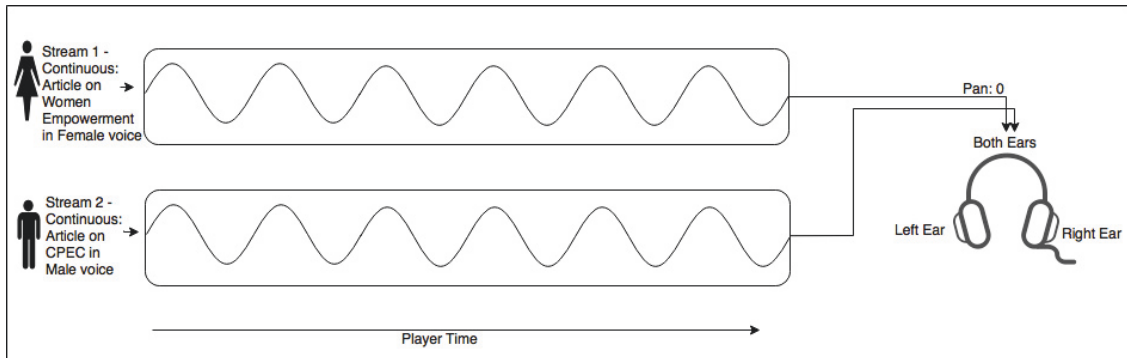


Figure 4.1 : **Continuous Stimulus Design for Study 2-A:** Presented streams are continuous and presented in both ears(panning parameter 0).

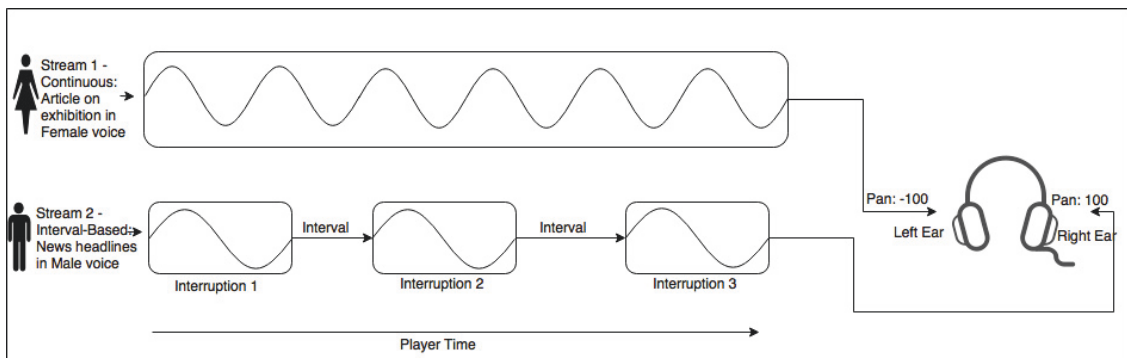
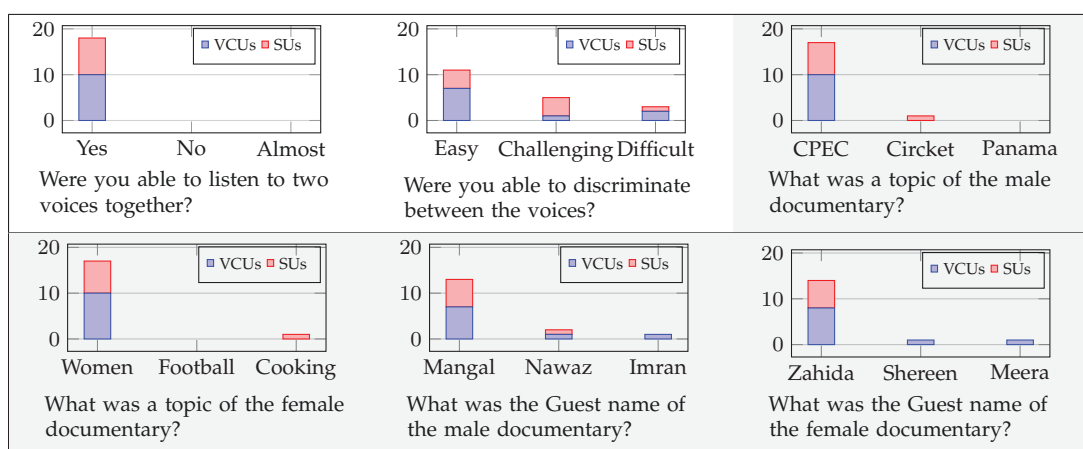


Figure 4.2 : **Continuous and Intermittent Stimulus Design for Study 2-B:** One stream is continuous, presented in the left ear (panning parameter -100), and the other stream is intermittent, presented in right ear(panning parameter 100).

Questionnaire I -

In Table 4.3, the experiential and basic content-based questions that were asked in Study 2-A are detailed. The number of correct and incorrect answers by the participants are graphically displayed.

Table 4.3 : **Questionnaire Study 2-A:** Questions and responses by users for Study 2-A. The questions in white background are experiential whereas the gray backgrounded are basic content-based questions. The first option in each content-based question is the correct answer.



4.2.3 Study 2-B - Continuous and Intermittent: Stimulus & Questionnaire:

The study 2-B incorporated the same method and stimulus, (*listen here* — — — <http://bit.ly/2RQaaTs>), of communicating two streams concurrently that was used in the study 1 discussed in chapter 3. In this study, an exhibition article and three headlines in the Urdu language were chosen. The exhibition documentary of stream 1 was played in the female voice of BBC Urdu's (BBC, 2016) notable newscaster 'Aaliya Nazki', whereas news headlines of stream 2 were in a male voice of BBC Urdu's famous newscaster 'Shafi Taqqi Jami'.

Based on dichotic listening, stream 1 was set to play continuously in the left earphone whereas stream 2 was played in the right earphone. Stream 2 played a silent interval of 20 seconds between each news headline. The total length of the

stimulus was 70 seconds. The stimulus design for study 2-B is illustrated in Figure 4.2.

Questionnaire II -

The questionnaire prepared for this study is shown in Table 4.4. Basic questions relate to prominent information that was asked, for example, what was the topic of the documentary, what was news headline indicating. In the advanced section, nine questions were asked based on less prominent information. The questionnaire also included questions that helped to measure user interest in concurrent multiple information communication.

4.2.4 Protocol

In these studies, the user did not have access to the audio player controls, such as volume, playback-rate, forward and back. An Apple MacBook Pro with left and right built-in audio speakers was used to play the prototype. Besides the built-in speakers, users were also provided with iPhone-6 earphones to listen to the audio streams. Users were first asked whether they were comfortable in using earphones or not, particularly the VCUs and given the choice to use any. The users used earphones to listen to the streams. No prior training was provided, however, users were orally briefed about information presentation / stimulus designs as illustrated in Figures 4.1 & 4.2. Users were provided with an idea through an example or two about the types of questions they would be expected to answer, for example, MCQs (closes), about the content they would hear. The questions were asked in an interview form in order to gain detailed responses to open questions set in the questionnaire. Users were told to focus on both the voice streams.

Table 4.4 : **Questionnaire Study 2-B: Questions (User Experience & Basic Content-based)** and responses by users for Study 2-B. The questions with white background indicate they are experiential whereas the questions with grey background are basic content-based questions. The dark grey background indicates they are advanced content-based questions. The first response in each content-based question is the correct answer.



4.3 Results & Analysis

The results of users response against each question in both the studies are indicated in Table 4.4. The analyses on results are described in subsequent subsections.

4.3.1 Qualitative Analysis

This subsection discusses the responses of the users qualitatively in order to share their comprehension and experience based on the non-structured interview conducted with them. Regarding users' experience, reactions, and expectations, the following points concisely discuss the factors reported by users based on their experience with such communication techniques. These factors provide several hints and open avenues for researchers to explore the directions of communicating multiple information concurrently.

Continuous vs. Intermittent Voice streams On the forms of deliveries, Users 4, 9, 12 and 18 (for profile details see table 4.1) reported that they were more comfortable in listening to continuous voices compared to intermittent communication. They mentioned that continuous voices with dichotic presentation could have been more helpful. Regarding intermittent communication, users reported that the volume of the secondary voice presenting intermittently broke their focus.

In contrast, User 1 found intermittent communication helpful. User 1 shared a comprehension technique that helped them to score well in the studies. They advised that in the dichotic listening stream, one needs to focus on the continuous voice and the mind would automatically catch the intermittent voice stream.

Language of the Content Highlighting the impact of the language of the content, user 11 reported that had the content been in their mother tongue, it would have helped them perform better in these studies.

Dichotic Presentation Dichotic presentation appeared as an important factor to differentiate the streams' content as it was reported by almost all the users. Only User 8 argued against the dichotic presentation and justified it by sharing that dichotic presentation created a focus shift issue. They added, human minds are used to listening to voices in both ears (diotic), but in dichotic presentation, the voices were split and entering in separate ears. Therefore, the brain started to capture information randomly sometimes from the right ear and sometimes from the left ear. Hence, User 8 argued that both voices should come to both ears because it is more natural than dichotic presentation.

Play Controls The provisioning of the audio player controls appeared as an essential demand by some users. User 13, 14, and 16 stated that if one wants to play multiple sounds concurrently then give control to the users so that they could set the value of controls according to their needs. For example, users should be able to bring one stream's volume low and others high or vice-versa. It was also mentioned that there was a need to adjust the playback-rate of the streams.

Interest in the Content Some users reported that their interest in the played content was a factor in comprehending information. User 15 reported that they could have focused more if the audio recordings were related to their interests, such as music or songs. Similarly, User 2 told that their interest in News helped to score better than had the stream not been of interest to them. Otherwise, they complained that they were unable to focus on both of the voices when they were played concurrently.

Keywords The results demonstrated that the keywords in the content contributed to the users being able to answer the questions correctly.

Training & Practice Multiple communication provides maximum results in minimum time but at the cost of losing some of the content. A user reported that it is not that easy to comprehend both voices together. The retention of content in memory is relatively lower than the sequential information. Users 6, 7 and 10 were of the view that practising on such a system can improve comprehension. User 10 also reported that in the study 2-A it was unexpected behaviour for the mind but by the end of the presentation, the mind helped to segregate the voices easily.

4.3.2 Quantitative Analysis

Besides stating the qualitative response of the users, quantitative analysis was also performed over users' responses based on user groups established from the user's academic profile, as mentioned in Table 4.2. This was performed in order to view the results from different perspectives. Group 1 (G1) include those users who possess at least a tertiary qualification. Those who were not holding tertiary education were placed in Group 2 (G2). It appeared that G1 performed better than the G2. Figure 4.3 shows percentages of correct and incorrect answers for each group.

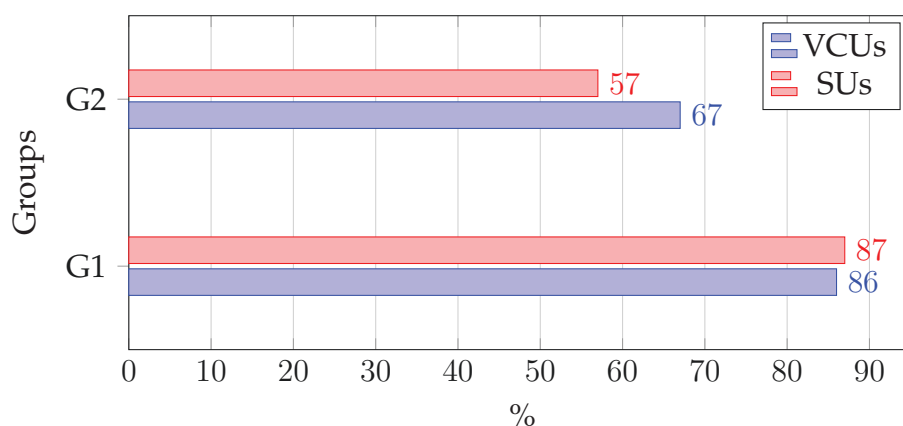


Figure 4.3 : **Group-wise Responses in Both Studies:** The percentage of correct answers to the questions asked in both the studies.

The performance of VCUs and SUs (sighted users) in comprehending infor-

mation was more or less the same as reflected in Figure 4.3. In this investigation, the users who did not have a tertiary qualification, but did have an interest in the news and technology, did well in answering the questions correctly.

The analysis also determined the performance of groups in answering both basic and advanced questions individually. The results are depicted in Figure 4.4 which indicates a similar pattern as seen in the Figure 4.3.

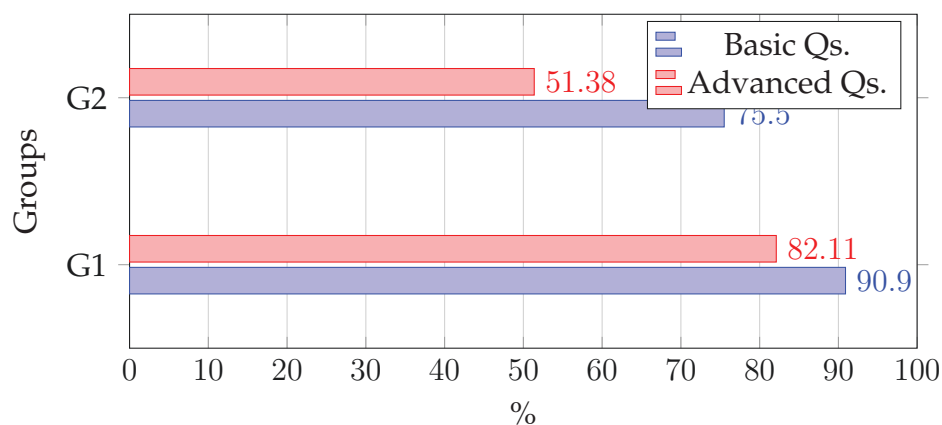


Figure 4.4 : **Group-wise Responses in Basic and Advanced Questions:** The percentage of correct answers to the basic and advanced questions.

A comparison of user's performance in both studies, to find which method, continuous or intermittent was more useful in communicating multiple content was performed. The analysis validated what was stated by many of the users; that the continuous content delivery was more appropriate than the intermittent communication. As shown in Figure 4.5, the percentage of correct answers in study 2-A were greater than the percentage in the study 2-B.

4.4 Discussion

To the question, 'whether you would prefer multiple information communication concurrently over the sequential flow of information?', there was an equal number of users that answered 'Yes,' 'No,' and 'Maybe'. Many of the users who answered 'No' argued that in the concurrent form of delivery, they might miss a significant

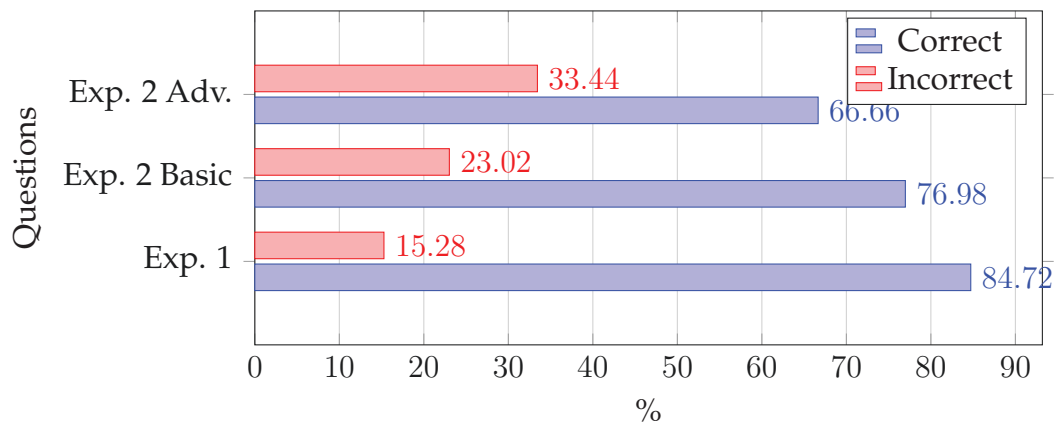


Figure 4.5 : **All Users' Responses in Both Studies:** The percentage of correct answers to the questions asked in both the studies.

amount of information that could be a big problem when the information requirement is crucial and requires listening to it carefully and uninterrupted. Therefore, users asked to provide them with authority regarding player controls to decide themselves whether they want two voices streams to be played concurrently or sequentially. Those who chose 'maybe' also argued that it would depend on the information-seeking context. Those who answered 'yes' mentioned the delivery of multiple information at the same time is the reason for opting this option.

The studies and response of the participants are encouraging to explore this avenue further. Although some of the users manifested disinclination in multiple voice information delivery, their score in the questionnaire reveals that they performed well in comprehending the information from voice streams played concurrently. Many of the participants expressed interest in the development of new technologies that may assist them in meeting the daily challenges of extensive information.

The approach taken in the study 2-B imitated a GUI overlay technique. The short audio clips of useful information were passed to the user while listening to a documentary by using almost half of the audio display bandwidth, as prototype played the headlines in the right ear that otherwise remained silent throughout

the clip. It was expected that this approach would be more acceptable to users. Contrary to expectations, most of the participants found intermittent communication a hindrance in comprehending multiple voice streams, as reflected in the above analysis. The non-optimal utilisation of auditory bandwidth and volume for auxiliary information is one possible reason for this hindrance. Therefore, the identification of optimal auditory bandwidth could be a subject of investigation to play the audio overlay.

However, in study 2-B, the dichotic listening based on audio panning (spatial separation) through simple balance technique helped users in segregating both the streams from each other. Almost, all the users agreed that 'panning' was helpful in separating the content from each other. Therefore, another approach for investigation could be delivering continuous streams (Study 2-A) dichotically, anticipating that it would increase the comprehension.

4.5 Vinfomize Framework

Based on the feedback received from both the sighted and the visually challenged users, the *Vinfomize* framework is developed. *Vinfomize* (V = Voice-based, info = Information, mize = optimization) is a framework for communicating multiple information concurrently. Many of the users opined that it is depended on the context, as to whether they would listen to multiple voice-based information concurrently or choose a sequential form of communication. For example, the students with a short time to prepare for the exams may choose a concurrent form of communication to revise the key concepts that they know already. Alternatively, if students do not have any particular time constraints, they might prefer to listen to a set of lectures sequentially for deep learning and understanding. Such information needs imply that the context of listening is important, and therefore that control should rest with the users.

These studies noted that multiple information communication depends on the

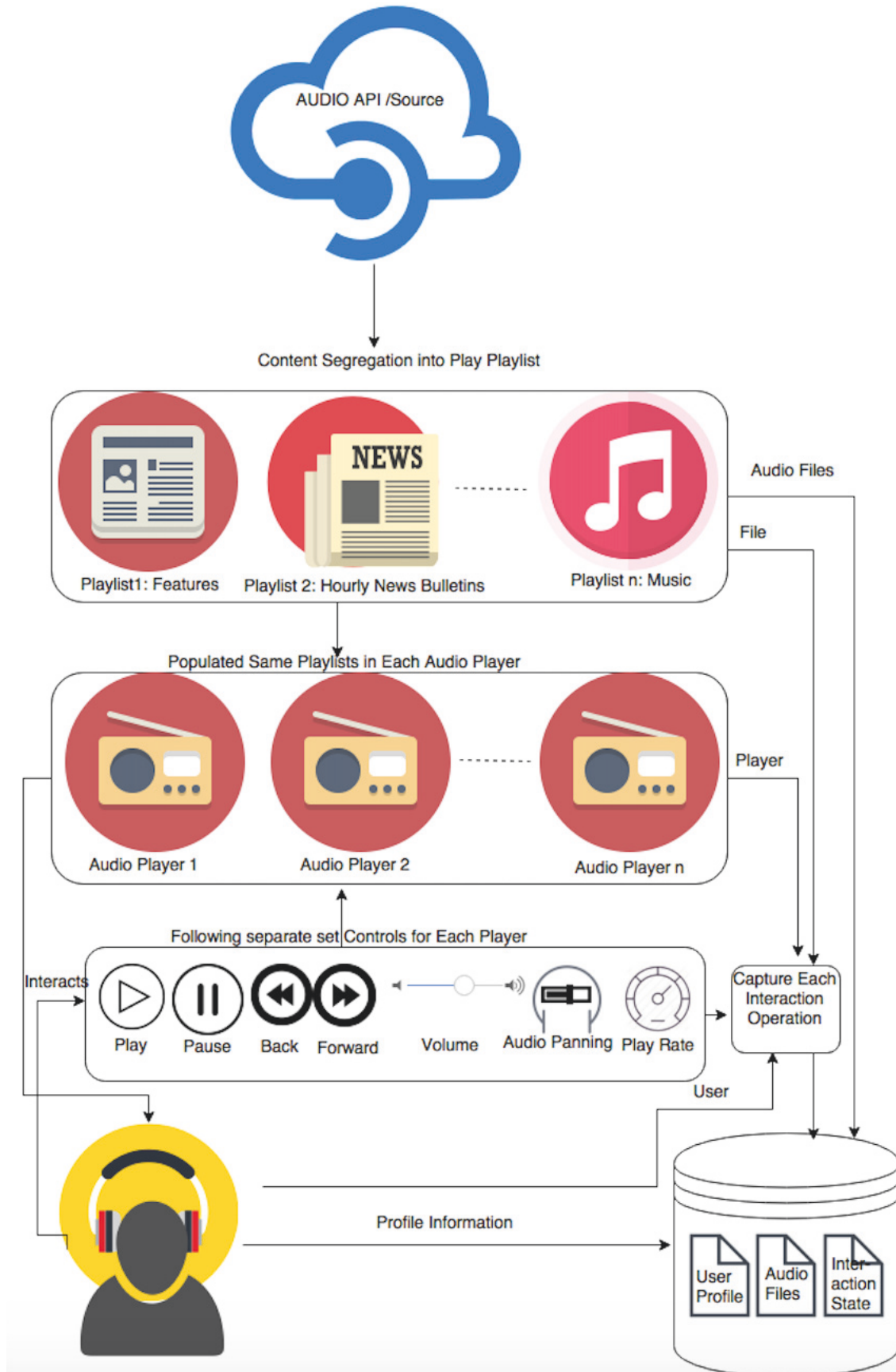


Figure 4.6 : **Vinfomize Framework:** A framework for optimising multiple voice-based information communication.

following factors:

- User auditory perception capabilities
- User information needs
- Information type
- Time constraint
- Physical context of the user

In *Vinfomize*, as illustrated in Figure 4.6, the concept of presenting multiple audio players with individually associated controls on a single page is introduced. Using these controls, users may play multiple players concurrently, or they can play a single player at any time. According to Guerreiro (2013), 2-3 concurrent streams renders better results for information scanning. *Vinfomize* enables users to set the audio controls according to their needs and receive optimal information from the system.

Each component mentioned in the *Vinfomize* framework is briefly discussed below:

Audio API

Audio API could be any source that provides various types of audio streams/files that the user may fetch.

Playlists

Playlists are organised lists that a user is able to create in order to categorise different types of content. Playlists might include news, features, talk shows, songs, sports commentary, drama, to name a few.

Audio Players

As mentioned already, a single web-page would have more than one audio player. Users may select an audio stream from their playlist using an audio player. Users can simultaneously play from more than one playlist using the equal number of audio players with the help of control sets associated with each player. For example, a user may play news on audio player 1, songs on the audio player 2 and live commentary on the audio player 3 by applying suitable settings on each of the players using audio controls.

Set of Controls

Each of the audio control included in *Vinfomize's* control panel is explained below:

Play / Pause A control button is provided to play and pause the audio stream on the relevant player. Interacting with this control on one player will not affect the other players.

Next / Previous Content Forward and back buttons are provided in *Vinfomize* for a user to be able to skip to the next track or go back to play a previous track in the playlist. Besides these buttons, a transport control bar is also provided that users may use to jump to the particular content within the audio stream.

Volume Many participants asked for volume control as they believed it would help them to segregate multiple streams. Using the volume control users may set the sound pressure level of the associated player. Setting different sound pressure levels for different players may help the users better to segregate the multiple streams from each other.

Panning Panning is a control that has been provided to induce the spatial difference between the competing streams. In this study, spatial difference or dichotic

listening was found to be one of the most critical factors for segregating competing streams. This panning control has been introduced to assist users to separate and identify the sources of the streams and therefore eventually help users to discriminate the streams from each other.

Setting the extreme values on the panning control would enable users to achieve dichotic listening. For example, for directing the two audio streams to separate ears, the user may set a pan control value to 1 for one stream and -1 for other. This setting would lead to playing the two streams together in separate earphones, the first one in the right, and the second one in the left.

Panning can be extended to the n number of streams. For example, a 3rd stream pan value can be set to 0 which would give diotic effect to the users which means the 3rd stream would be played in both ears and this effect would give an impression that the stream is being played at the 90 degree angle of the frontal horizontal plane, while the other two streams might be entering the left and the right ear separately. Similarly, users may set any value from 1 to -1 for each of the streams, and the values set by the users would determine the angle of the presentation of the stream to the users.

Playback-Rate In this experiment, multiple users commented that the speed of information delivery hindered them to grasp the information. Therefore, using the playback-rate control would allow users to be able to set the speed according to their preference.

We presented playback-rate control in the framework to increase or decrease the audio playback speed while maintaining the the absolute pitch of the reproduced audio content. Playback-rate can help users to quickly listen to the information, particularly in the sequential form of communication. The value of this control can be set from 0.5 to 2 to adjust the playing speed of the streams. This control may also be used in concurrent listening, as Guerreiro (2013) reported that the concurrent

speech with slightly higher playback-rate enables a significantly quicker scanning for relevant content.

Users The most important part of the framework interaction is the user. Though this framework is applicable on system speakers, it is recommended that the user should use earphones as it would not only provide the opportunity for a greater panning effect, but it may also minimise the masking from the user's surroundings.

Interaction & Database

Since communicating concurrent information is in the development phase, a database is proposed to store all user interactions for post-analysis. Since this component is added in the framework for the purpose of further research and enhancements, the designers may opt not to use this feature of the framework.

4.6 Web-based System Development based on the proposed Framework

Following the proposed framework, a web-based system capable of communicating concurrent information streams with the help of discussed audio controls is developed. This system is developed using the World Wide Web Consortium (W3C) *web audio API*, that besides the conventional controls, also provides the flexibility to include panning and high playback-rate controls. In the prototype system, multiple audio players, each provided with a populated playlist are presented. Users are able to select a audio stream from the playlist on one player, and another information stream on the second audio player using the audio controls provided in the system.

It is expected that a large number of users with distinctive profiles from different walks of life would interact with the prototype when presumably be launched,

for further research after this thesis. *Users' In-the-Wild Usage* (Guerreiro, 2016b) interactions will be recorded and documented within the system. A comprehensive analysis will be performed on the user interactions with the system. This analysis will expectantly give designers an in-depth knowledge on how users have interacted with the system and what improvements can be introduced in information delivery on the basis of this in-depth knowledge.

The screenshot of the system interface is displayed in Appendix-B. The functional system is accessible at the following URL:

<http://www.utsresearch.com/vinfomize>

(p.s. If URL is inaccessible, you may find it by writing search keywords: 'Vinfomize concurrent communication' in a search engine.)

4.7 Limitations & Future Work

This was an informal study primarily aimed that involve VCU and SU and to determine whether participants are interested in concurrent information communication; and by what type of interaction approach users wish to interact with such systems. The investigation was of an exploratory nature, and user's participation was casual which helped us to obtain their free perspectives. Their valuable feedback helped us to propose a framework and also build a prototype, but the study barely carried a standard to make a research claim. The study is extended by taking user's feedback into account and designed standardised experiments that are discussed in the following chapters with the following aims:

- to compare the high playback-rate, and the basic concurrent design to determine which is more effective in communicating information to users and which remains close to the information transfer efficiency that is achieved in conventional sequential speech-based information communication
- to examine designs for speech communication that can communicate concur-

rent speech-based information equal to the information transfer efficiency that is achieved in conventional sequential speech-based information communication

- to evaluate comprehension depth of both the primary and the secondary information streams in speech-based concurrent information designs and determine which design remains the most effective in communicating speech-based information concurrently

The investigation regarding each aim is separately discussed in the subsequent three chapters.

An earlier version of the research discussed in this chapter has been published in the following papers:

Publication 2: M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, “Concurrent Voice-Based Multiple Information Communication: A Study Report of Profile-Based Users’ Interaction,” in *145th Convention of the Audio Engineering Society*. Audio Engineering Society, 2018. Attached as Appendix-H.

Publication 3: M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, “Vinfomize: A framework for multiple voice-based information communication,” in *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*. ACM, 2019, pp. 143–147. Attached as Appendix-I.

Chapter 5

A Comparison between High Playback-rate and Concurrent Design

In the previous experiment, participants mentioned that the spatial separation between the concurrent sources helped them in segregating the streams, and also a few users considered that the intermittent form of concurrent communication was useful to them. Therefore, based on the literature review, results and feedback received from the users in previous studies, research investigation regarding speech-based concurrent information is extended and the standardised experiment was carried out to explore the following three broad aims:

- Compare a basic concurrent design to a doubled playback rate design, to determine which is more effective in communicating information to the users and which remains close to the information transfer efficiency that is achieved in baseline sequential speech-based information communication
- Examine designs for speech communication that can communicate concurrent speech-based information equal to the information transfer efficiency that is achieved in conventional baseline speech-based information communication
- Evaluate comprehension depth of both the primary and the secondary information streams in speech-based concurrent information designs and determine which design remains the most effective in communicating speech-based information concurrently

These aims are met by designing a comprehensive experiment methodology which is discussed in this chapter. Thirty four users participated in this experiment

that provided a comprehensive data enabling us to investigate each aim. The investigation regarding each aim is individually discussed in this and subsequent two chapters (6 and 7).

This chapter reports on the first aim of investigating a high playback-rate design and a basic concurrent design (The other concurrent designs that were tested in the same experiment are discussed in detail in the subsequent two chapters). In the high-playback-rate design, multiple information streams were communicated sequentially by doubling the normal playback-rate, while in the concurrent design, two speech-based information streams were played concurrently without involving panning (spatial difference). Both these designs were compared with a benchmark design set from the baseline condition where two streams were sequentially played with regular(1x) playback-rate.

5.1 Aims & Motivation

Aims: The fundamental aim is to investigate the possibilities of communicating multiple-speech based information streams efficiently. The analysis in this chapter set out to satisfy the following questions: a) How different the comprehension appears for concurrent and the high playback-rate designs when compared to the baseline condition? b) Do both the high-playback rate and concurrent playback design render similar comprehension? c) Does the comprehension pattern in all these designs remain the same?

Motivation: The motivation for this investigation is to fulfil users' needs to seek information quickly and efficiently while interacting with the system. Based on the results of this investigation, appropriate information communication approaches can then be designed which help to communicate more information to listeners, and thereby increase the efficiency of the information communication. The quick communication of multiple information streams can be helpful in listening to

digital streams, relevance scanning, scanning for specific information, notifications using a secondary audio channel and navigation. This study can also help to inform the design of complex and information-heavy speech interaction methods.

5.2 Method

The experiment investigating above three aims is outlined below. In chapters 6 and 7, which discuss the data analysis focusing on the remaining two aims, the method section is not repeated.

5.2.1 Participants

After receiving institutional Human Research Ethics Committee approval for the research protocol (attached as Appendix-C), user participation campaigns were launched. Participants were selected based on two criteria: 1) not having a significant hearing impairment, and 2) having competent English language skills, as the listening experiment's content was in the English language. In total, 34 participants, 14 female and 20 male, took part in the experiment after providing consent. The mean age of the participants was 26 with a standard deviation of 6.

5.2.2 Design

Concurrent Condition

Independent Variables We manipulated three independent variables within the designs:

Content Type: The stimuli were either created from mono-channelled DCT (Discourse Comprehension Test) or the stereo-channelled IELTS (International English Language Testing System) audio content files.

Presentation Form: The presentation form was either concurrent with two continuous streams or concurrent with one continuous stream and another

intermittent.

Spatial Configuration: The spatial configuration that the stimuli were presented in was one of three options: Diotic, Diotic-Monotic or Dichotic.

Within the concurrent condition six distinct stimuli designs were devised to communicate two speech-based information streams on separate topics concurrently. One stream was in the female (higher fundamental frequency) voice, and the other was in the male (lower fundamental frequency) voice. From the six stimuli designs, three designs followed the first form, and the remaining three followed the second form of communication from the following list:

- Continuous Female Stream with Continuous Male Stream (Continuous)
- Continuous Female Stream with Intermittent Male Stream (Intermittent)

Each of the Continuous and Intermittent based stimuli design was individually applied with one of the following three pan conditions to involve a spatial difference in streams presenting streams to the specific ear(s):

- 0,0 – Diotic (Both Streams in **both ears**)
- 0,100 – Diotic-Monotic (Female stream in **both ears** whereas the Male stream in the **right ear**)
- -100,100 – Dichotic (Female stream in the **left ear** whereas the Male voice stream in the **right ear**)

All the six design methods were repeated on two types of audio content material that increased concurrent stimuli designs to 12. The audio types of content material were:

- Discourse Comprehension Test (DCT)
- International English Language Testing System (IELTS)

Each of the rendered concurrent stimuli design is described in Table 5.1 and illustrated in Figure 5.1 for further clarity.

Table 5.1 : **Speech-based Concurrent Communication Designs**: The name of each design is mentioned in the first group comprising first three columns along with the attributes of primary and the secondary streams in separate groups comprising three columns each.

Concurrent Design			Primary Stream			Secondary Stream		
Content	Form	Pan	Voice	Presen.	Ear	Voice	Presen.	Ear
DCT								
DCT	Conti.	Diotic	Female	Conti.	Both	Male	Conti.	Both
DCT	Conti.	Dio-Mon	Female	Conti.	Both	Male	Conti.	Right
DCT	Conti.	Dichotic	Female	Conti.	Left	Male	Conti.	Right
DCT	Intermi.	Diotic	Female	Conti.	Both	Male	Intermi.	Both
DCT	Intermi.	Dio-Mon	Female	Conti.	Both	Male	Intermi.	Right
DCT	Intermi.	Dichotic	Female	Conti.	Left	Male	Intermi.	Right
IELTS								
IELTS	Conti.	Diotic	Female	Conti.	Both	Male	Conti.	Both
IELTS	Conti.	Dio-Mon	Female	Conti.	Both	Male	Conti.	Right
IELTS	Conti.	Dichotic	Female	Conti.	Left	Male	Conti.	Right
IELTS	Intermi.	Diotic	Female	Conti.	Both	Male	Intermi.	Both
IELTS	Intermi.	Dio-Mon	Female	Conti.	Both	Male	Intermi.	Right
IELTS	Intermi.	Dichotic	Female	Conti.	Left	Male	Intermi.	Right

Baseline Condition

Under this condition, a baseline stimulus representing the conventional speech-based communication was designed where the continuous female information stream followed by a continuous male information stream was presented sequentially without involving spatial difference, concurrency, or any increase in playback rate. The purpose of this design was to determine a benchmark of user comprehension in the baseline condition that could subsequently be used to evaluate users' comprehension in other conditions.

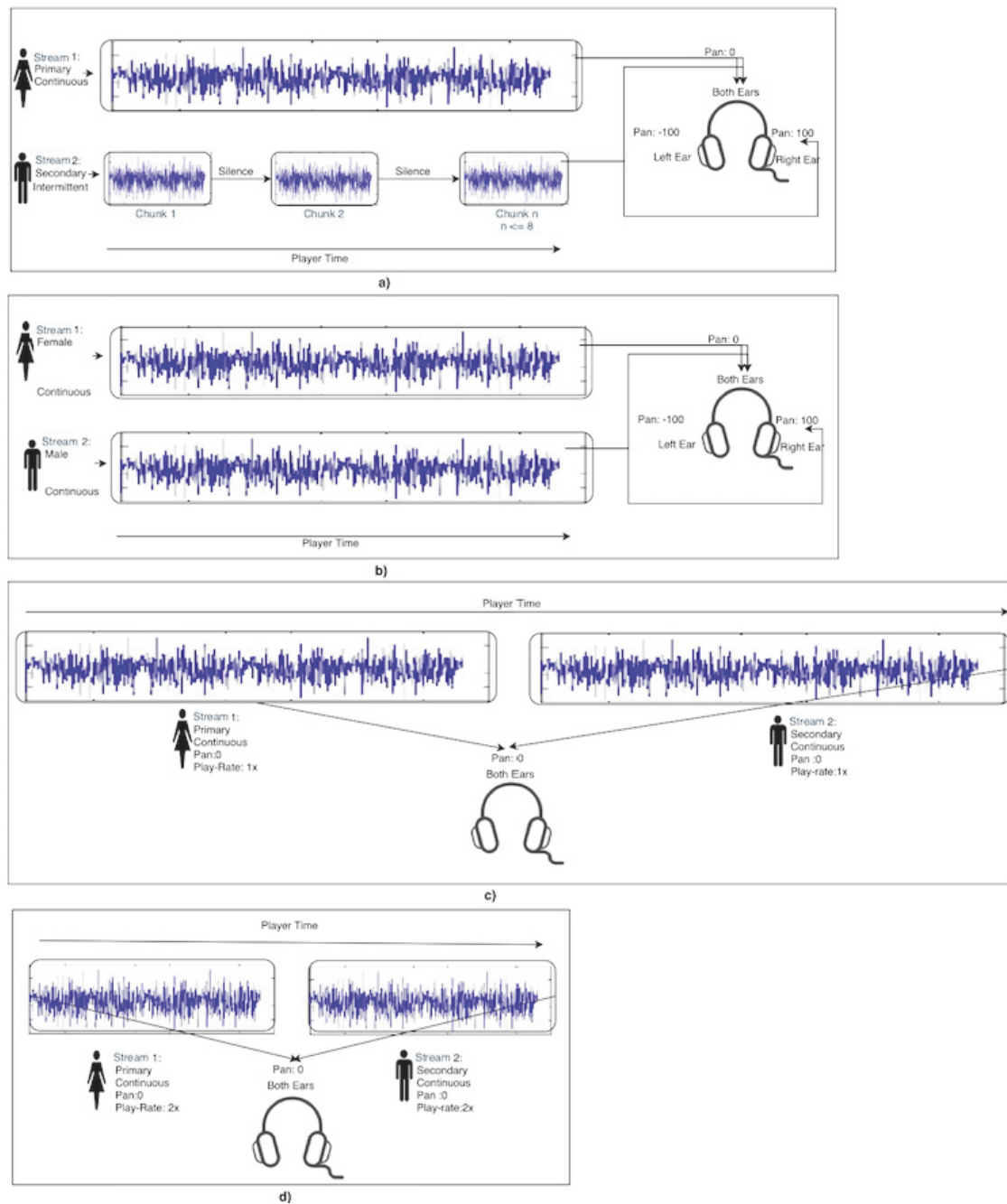


Figure 5.1 : **Stimuli Designs:** *a)* Intermittent: Continuous Stream in Female Voice with Intermittent Stream in Male Voice having Panning / Spatial Conditions of 0-0 (Diotic), 0-100 (Diotic-Monotic), (-)100-100 (Dichotic), *b)* Continuous: Continuous Stream in Female Voice with Continuous Stream in Male Voice having Panning / Spatial Conditions of 0-0 (Diotic), 0-100 (Diotic-Monotic), (-)100-100 (Dichotic), *c)* Baseline: Continuous Stream in Female Voice followed by Continuous Stream in Male Voice with no Panning Conditions (0), *d)* Doubled Playback-rate: Continuous Stream in Female Voice followed by Continuous Stream in Male Voice with doubled play-rate

Doubled Playback-rate (Seq-2x) Condition

Besides the baseline stimulus, another sequential stimulus (Seq-2x) was also designed where streams were played following the baseline stimulus method where the only difference in playback-rate that it was doubled (2x). The purpose of this design was to test the communication of multiple information in unit time as shown in Fig. 5.1-d.

5.2.3 Material

For speech-based stimuli designs, two types of content resources were used Discourse Comprehension Test (DCT) and the International English Language Testing System (IELTS).

Discourse Comprehension Test (DCT)

The commercially available Discourse Comprehension Test (DCT) (Iyer et al., 2013; Obermeyer and Edmonds, 2018; Welland et al., 2002) is a standardised test that primarily assesses the comprehension and retention of spoken narrative discourse by adults suffering from aphasia. The test contains 12 stories where each story had a length from 73 to 95 seconds which describes a humorous situation. The material purchased from (DCT) was received on a CD having twelve mono-channelled audio tracks each presenting a story in the male (lower fundamental frequency) voice. Each track was exported into .wav format with the sample rate of 44.1KHz and the bit rate of 16 using the Apple iTunes software. Since the conceived stimuli designs were planned to be discriminable by gender (fundamental frequency) i.e. female (higher frequency) voice and male (lower frequency) voice, therefore, the pitch of the six from twelve tracks was changed by increasing it 17% from the default lower frequency male voice using an open source software (Audacity). This increase in pitch transformed the lower frequency voice into a higher frequency voice and was applied to six stories, while six others were left in the original male

voice.

International English Language Testing System (IELTS)

The IELTS listening material was also used in the experiment as it was readily available and provided heterogeneous content in stereo-channelled audio files. For the experiment, 12 audio files containing monologue audio content were selected (sample rate 44.1KHz, 16-bit). Six files in the male (lower fundamental frequency) voice and a further six in the female (higher fundamental frequency) voice were selected. From each monologue file, an initial 58-70 seconds of the meaningful content was extracted.

5.2.4 Stimuli Information

In total, 24 Continuous speech-based streams were obtained and processed from both types of material. For having the Intermittent streams, the contents of the half of the Continuous stream in male voice were broken into temporal segments (chunks) by giving silent intervals of 5 to 10 seconds in them. Each stream was repeatedly applied with each of the three pan conditions 0, 100 and -100 that rendered 72 (24×3) streams where 36 were in the female voice, and 36 (18 Continuous and 18 Intermittent) were in the male voice. Then each of the rendered male streams was repeatedly combined with the female stream of the same material using the Audacity software for Mac. This multiplication generated 216 combinations to incorporate randomisation in the experiment for minimising the combinational effect in the analysis. From 216, 12 stimuli (6 DCT + 6 IELTS) were selected randomly (without replacement) to be presented to each user, where each stimulus was a case of one of the designs mentioned in table 5.1. The length of each rendered stimulus was within 55 to 90 seconds except the baseline design. Besides the 12 concurrent designs, the additional two designs, baseline and Seq-2x, were presented to participants.

5.2.5 Measures

After listening to each stimulus design, participants answered the questions, discussed in section 5.2.6, from the stimulus. Since each stimulus was the combination of two streams and each stream had a set of 8 questions, a user answered a total of 224 questions with yes/no/don't options. The user comprehension was measured on the basis of the number of correct answers after listening to each stimulus.

In previous experiments by the authors, discussed in chapters 3 and 4, users often pointed out that they did not know the answer and were looking to select a 'Don't know' option, which was not present, and therefore were compelled to choose a probable answer from the given options. This necessarily results in less accurate estimations of user comprehension of the stimulus content, with the assumption being that participants who don't know the answer will naturally choose one of the remaining two options equally, thereby increasing noise. Therefore, in this experimental protocol a third option, 'Don't know' was included, in addition to the usual 'Yes' and 'No' user responses, in order to carefully distinguish these cases.

5.2.6 Questionnaire

The DCT material was accompanied with the default questions that were used in the experiment as is, however, for IELTS new questions following the DCT pattern were prepared. Each story had eight questions having yes/no/don't know answers. The questions were arranged into assessment categories to assess the depth of comprehension by the users. For each following category type, two questions were arranged:

- Main Information Stated (MIS)
- Main Information Implied (MII)
- Detailed Information Stated (DTS)

- Detailed Information Implied (DTI)

The questions in MIS were constructed from the main stated information of the story. These questions assessed how much a participant had comprehended the main idea that was repeated or elaborated by other information in the story (main information). The MII questions were based on the information that was not directly discussed in the story, but a user had to infer it from the stated main information. The questions in DTS were framed from the stated information of the story that estimated the comprehension of detailed information. Detailed information was mentioned only once and not elaborated by other information in the story. DTI questions were based on the information that was not directly explained in the story, but a user had to infer from the detailed information. The implied questions examined whether a user was able to make a mental map or bridging assumptions of the information or not.

5.2.7 Apparatus

To minimise the participation time for completing the tests and for convenience, a web-based system using PHP, MySQL, JQuery, HTML5, CSS, and Bootstrap was designed to play the stimuli. The web system was accessible using a web browser where 14 HTML audio players, each playing one stimulus design, were presented on the screen along with the questions under the relevant stimulus player. The entire system including interface, the randomly selected 14 stimuli designs and the relevant questionnaire is shown as Appendix-D. The tests were conducted in a quiet purpose-built Creativity and Cognition Studios (CCS) of the University of Technology Sydney. Three identical Apple iMac computers having 2.7GHz quad-core Intel Core i5 processor, 8 GB RAM, installed with Yosemite 10.10.5 OS were arranged in the studio. To listen to the audio stimuli Beyerdynamic's DT770 250 OHM headphones were used that were connected to the headphone jack of the computer. Users were provided with the gain control only to set the intensity

as per listening choice. Since three computers were used in the studio, therefore, at a time, up to three participants engaged in the experiment simultaneously.

5.2.8 General Procedure

The selected users were verbally briefed on the study protocol before the start of the experiment, and the instructions were presented on a screen after registration. Before starting the experiment, users entered their demographic profile information that included, name, age, qualification, first language, country, hearing impairment and type of computer and headphones used in case of participating from outside of the CCS. At the end of the experiments, user's subjective response to the concurrent and sequential information communication was also obtained by asking three questions related to user experience. All users' responses were stored in the MySQL database for the post-experiment analysis.

The entire experiment interface including form obtaining user's demographic profile information, playable URLs of stimuli representing each design, and associated questions to the stimulus from the questionnaire is produced in Appendix-D.

5.3 Results

Targeting the first aim of this study, three aspects are covered. First, the proportion of users' selection of options for answering the questions in the baseline, high playback-rate, and the concurrent (IELTS.Continuous.Diotic) designs are evaluated in sub-section 5.3.1. After that, users' performance in comprehending the information regarding answering questions correctly in the same three designs is compared in sub-section 5.3.2. Thirdly, the depth of information comprehension is assessed for the same three designs in sub-section 5.3.3. In this analysis, the IELTS.Continuous.Diotic is selected from the concurrent designs for comparison because this design is the most basic concurrent design that does not involve any intervention, for example, panning and intermittence, except playing two audio

streams to both the ears concurrently.

5.3.1 Proportion Analysis

This section evaluates the proportion of selecting options for answering questions in all the three designs individually. In the baseline design, as shown in the figure 5.2, users frequently selected 'Don't Know' option for both types of the 'Yes' & 'No' expected answers. It showed that when the expected answer was 'No', 23% responses were selected 'Don't know' by the users which were significantly higher than 8% 'Don't Know' responses in the condition when the expected answer was 'Yes'. Also, The percentage, 18%, of selecting 'Yes' as a wrong answer was higher than the 14% of selecting 'No' as a wrong answer.

In higher-play-rate approach, as shown in the figure 5.2, the proportion of selecting 'Don't Know' was higher than the baseline conditions. In this approach, when the expected answer was 'No', the proportion of selecting Don't know was 29% and when the expected answer was 'Yes' the percentage was 25%. Moreover, opposing to the baseline condition, the percentage of selecting 'Yes', 18%, as a wrong answer was less than the percentage of selecting 'No', 22%, as a wrong answer.

Regarding the concurrent design, figure 5.2 shows the proportion of response submitted by the users. Similar trends appeared in this design as seen in the high playback-rate design. In this design, when the expected answer was 'No', the proportion of selecting Don't know was 35% and when the expected answer was 'Yes' the percentage remained 31%. Generally, the proportion of giving correct responses and selection of 'don't know' appeared similar as seen in the high playback-rate design.

Overall, in all the three approaches, users selected all three 'Yes,' 'No' and 'Don't Know' options as answers to the questions.

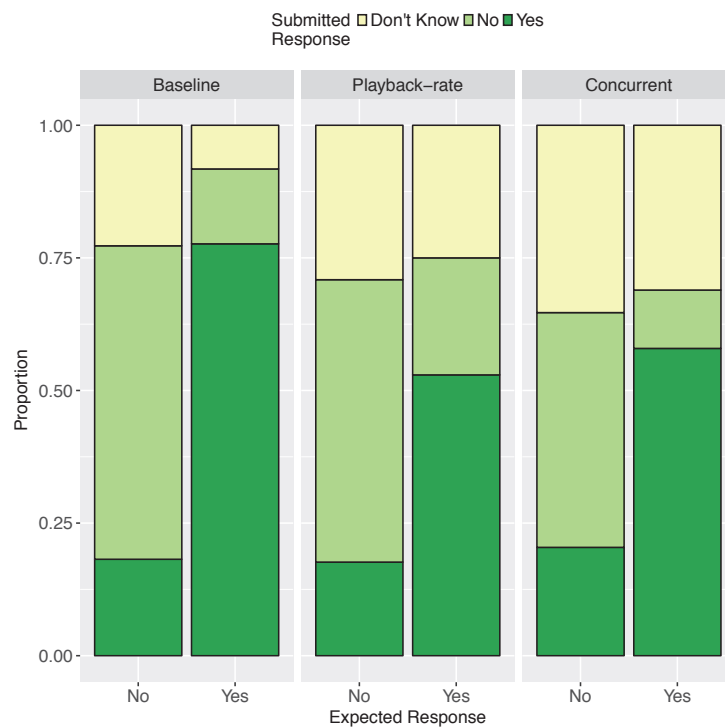


Figure 5.2 : **The Proportion of User Responses:** across the baseline, playback-rate and concurrent designs.

5.3.2 Comprehension Performance Analysis

In the second part of this analysis, the comprehension is assessed by calculating the percentage of the correct answers for all three designs. For the percentage calculation, users response matching to the expected answer was counted as a correct answer whereas the opposite answer or the selection of 'Don't Know' options were considered as the wrong answer. The assessments of all three designs are discussed individually.

In the baseline design analysis, the percentage of the correct answer was calculated to set a benchmark. As figure 5.3 shows, 63% questions were correctly answered by the users. Inversely, 37% questions either were answered incorrectly, or users did not know the answers implying that user could not fully comprehend the content to answer all the question correctly. The percentage, 63%, of giving the correct answer in the baseline sequential information communication set the

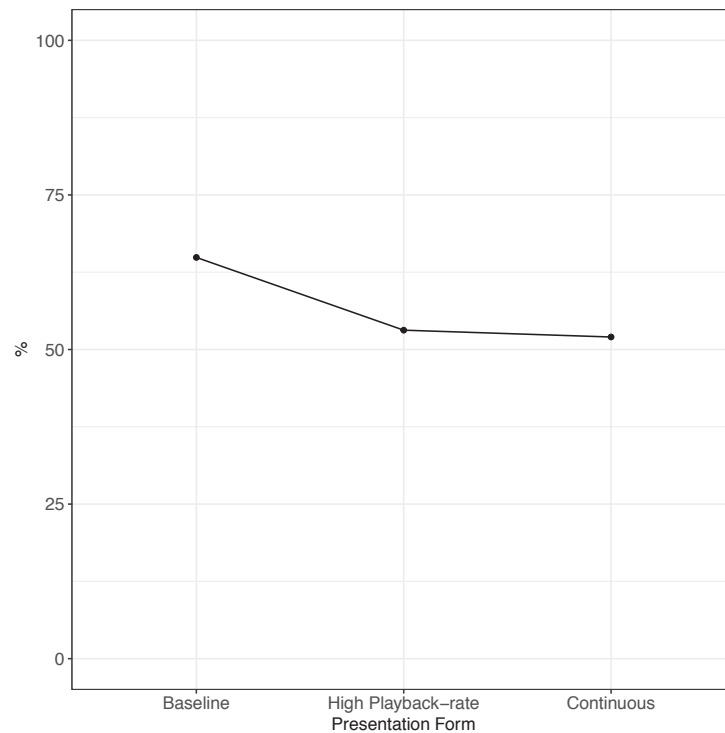


Figure 5.3 : **The percentage of correct answers:** across the baseline, high playback-rate and concurrent designs.

benchmark to compare users' comprehension in the high playback-rate and the concurrent designs.

Regarding the high playback-rate design, figure 5.3 shows the percentage of correct answers. In this approach, the percentage of giving correct answers was 53% that using the proportion test (Wilson, 1927; R, 1998) shows the users' comprehension performance was significantly lower ($p < 0.001$) than the baseline design.

Similarly in the concurrent design, as shown in the figure 5.3, the percentage of correct answers was 52%. This indicates that the users' comprehension in the concurrent designed was similar ($p 0.761$) to the higher-playback-rate design.

Overall, both the high playback-rate and the concurrent designs performed similarly in delivering multiple information. However, they were significantly lower ($p < 0.001$) than the benchmark set from the baseline approach.

5.3.3 Comprehension Depth Analysis

The third part of the analysis focuses on an evaluation of the comprehension depth of the content in the baseline design, which is compared to the high playback-rate and the concurrent designs.

In the baseline condition, the individual percentage for MIS, MII, DTS, and DTI is calculated to set a benchmark that was later used to compare the comprehension in the quick designs. Figure 5.4 using a red line shows the analysis of the Baseline design. The percentage of correct answers to the questions set from MIS was 85% whereas, in MII, DTS, DTI, it was 72%, 51%, 51% respectively. The analysis shows that the comprehension of MIS was significantly higher compared to the other information categories.

Following the pattern adopted in the analysis of baseline design, the comprehension depth was evaluated for the high playback-rate design. Figure 5.4 using a blue line shows the percentage of correct answers in the concurrent design with respect to MIS, MII, DTS, DTI. In this design, the percentage of correct answers to the questions set from MIS was 67% whereas in MII, DTS, DTI it was 51%, 58%, 36% respectively.

Similarly, in the concurrent design, the percentage of correct answers to the questions set from MIS was 67% whereas in MII, DTS, DTI it was 60%, 41%, 39% respectively. This shows that the pattern of information comprehension in this concurrent approach was similar to the comprehension depth calculated in the high playback-rate design, except one condition where the percentage regarding DTS in the high playback-rate design remained higher than the concurrent design. The comprehension assessment for the concurrent design is reflected using a green line in the figure 5.4.

Besides calculating the comprehension depth, each MIS, MII, DTS, DTI data point (percentage) in high playback-rate design and the concurrent design are also statistically compared with the relevant data points in the benchmark set from

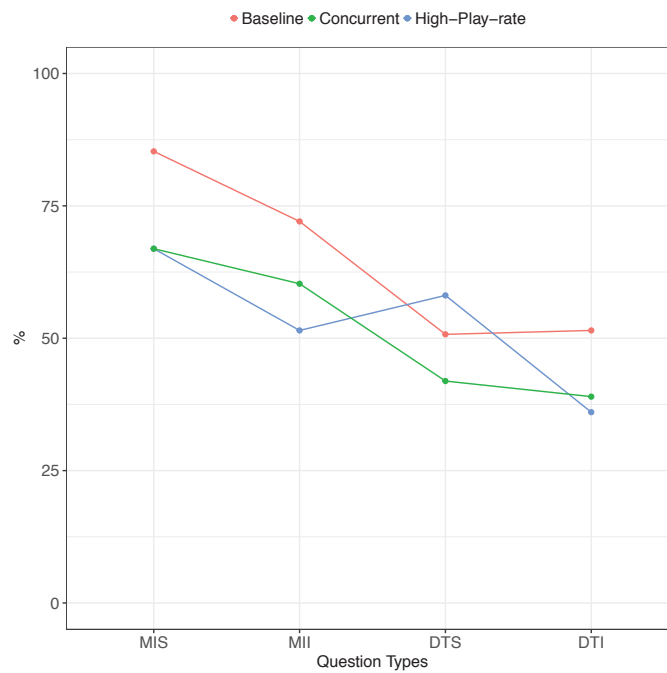


Figure 5.4 : **Comprehension Depth:** percentage of correct answers with respect to MIS, MII, DTS, DTI across the baseline, high playback-rate and concurrent designs.

the baseline condition and mentioned in Table 5.2. In almost all the data-points related to the main information, the comprehension in the quick designs was lower when compared to the baseline condition, whereas in the data-points related to the detailed information, the comprehension remained similar in all three designs.

Table 5.2 : Results of the one-to-one proportion comparison between the correct answers in the baseline and the other quick designs with respect to MIS, MII, DTS, DTI, with Bonferroni correction, *** < 0.001

Concurrent Design	Primary Stream			
	MIS	MII	DTS	DTI
High Playback-rate	***	***	0.273	0.014
Concurrent	***	0.054	0.181	0.051

5.4 Discussion

Compared to the baseline condition, users' comprehension dropped significantly in the high playback-rate and the concurrent designs. However, in both the designs users were able to answer more than 50% of the questions correctly. In the results analysis, both designs, high playback-rate and the concurrent, performed similarly in communicating information as users' comprehension score was almost the same. Moreover, in answering the questions set from the detailed category, the comprehension was similar in all three designs. User's ability to answer over 50% of questions shows that both approaches have the potential to be used for communicating multiple information streams quickly. Therefore, further investigations to explore more possibilities and come up with optimal designs for efficient multiple information communication.

Moreover, as discussed in the methods section, the questions were arranged in 4 categories MIS, MII, DTS, DTI formed by information repetition in the content to assess the comprehension depth. It was expected that the users content comprehension would remain in the same order mentioned above, as the main information was repeated multiple times in the content whereas the detailed information was played once only in the content. In all the three designs, users' comprehension was higher in MIS and MII and was lower in DTS and DTI, with the exception of comprehension in DTS in the high playback-rate design. This shows that in all three designs, the pattern of comprehension depth remained similar.

In comparing the viability of the efficient designs, both approaches rendered similar comprehensibility, but the concurrent design has certain advantages over the high playback-rate design. Some of the main advantages are:

- Supports live streaming - for example if two radio programs are being broadcast live at the same time, users would be able to listen to both of them at the same time with concurrent presentation. In other words, concurrency

may help users to listen to two different streams simultaneously.

- Selecting and attending an information stream - for example, if two streams are provided concurrently users using their selection and attention abilities may switch focus immediately towards the information stream that carries high user interest.
- Divided attention - Users may get the gist from both the streams at the same time using divided attention.

These advantages show that the concurrent approach can provide more efficient communication for the multiple streams to the users. Therefore, the exploration of the concurrent approach would be continued, and different designs from this study would be analysed to come up with an optimal design that could communicate multiple speech-based information streams optimally.

5.5 Limitations and Future Work

This study shows the potential of communicating multiple information using the high playback-rate approach and the concurrent approach. This study compares two designs, one from each approach, to communicate multiple information. There can be many configurations in each approach that can be tested for efficient communication of audio streams.

Some of the users had reported excessive cognitive load while listening to the quick approaches. It could be an exciting investigation to test more design configurations to identify how the cognitive load could be minimised while communicating multiple information streams quickly. In chapters 6 and 7, more concurrent designs are thoroughly investigated.

An earlier version of the research discussed in this chapter has been presented in the following paper:

Publication 4: —, “Investigating Efficient Speech-based Information Communication - A Comparison between the High-rate and the Concurrent Playback Designs,” *Journal on Multimodal User Interfaces (JMUI)*, vol. -, no. -, pp. 1–8, 2019, submitted.

Attached as Appendix-J.

Chapter 6

Investigating Concurrent Speech-based Designs for Information Communication

This chapter extends the analysis of the comprehensive experiment discussed in Chapter 5. The results from the previous chapters support the viability of the concurrent information communication. This chapter analyses various concurrent approaches that were designed in the experiment. The experiment was mainly designed to propose and test various designs for communicating speech-based information concurrently. For testing different concurrent designs, as mentioned in table 5.1, two audio streams from two types of content were played concurrently, in both a *continuous* or *intermittent* form, with the manipulation of a variety of spatial configurations (that is, *Diotic*, *Diotic-Monotic*, and *Dichotic*). In total, 12 concurrent speech-based design configurations were tested with each user. In this analysis, determining the effectiveness of each design in comprehending information by the users is aimed.

6.1 Aims & Motivation

Aims: The second aim of this study is to examine designs for speech communication that can communicate concurrent speech-based information similar to the information transfer efficiency that is achieved in conventional sequential speech-based information communication. Additionally, the analysis is performed to satisfy the following questions:

- a) Which concurrent form presentation, *continuous* or *intermittent*, provides better user's comprehension?
- b) Does the spatial difference between the concurrent

streams improve user's comprehension?

Motivation: If this study remains successful, concurrent speech-based communication designs that render better information communication can be adopted in speech-based interaction to communicate more information to listeners in an efficient manner and can help to guide the design of complex and information-heavy speech interaction methods. Concurrent speech can be helpful in listening to two TV streams, relevance scanning, scanning for specific information, notifications using a secondary audio channel, TV navigation, and subtitles and assisted navigation, to name a few.

6.2 Methodology

Discussed in section 5.2.

6.3 Results

The results analysis of this Chapter targets the second aim of the study, therefore, speech-based concurrent information designs are evaluated. For this, users' comprehension in all concurrent designs were measured and compared with their comprehension in the baseline design (benchmark). For each design, the analysis included two parts: 1) comparing the proportion of users' responses, and 2) calculating the percentage of correct answers. Afterwards, the intermittent presentation form, identified as the highest-producing comprehension in concurrent designs, was further investigated to assess the underlying mechanism and the users' comprehension behaviour. The results of concurrent designs analysis and the intermittent designs in detail are individually discussed in following sub-sections.

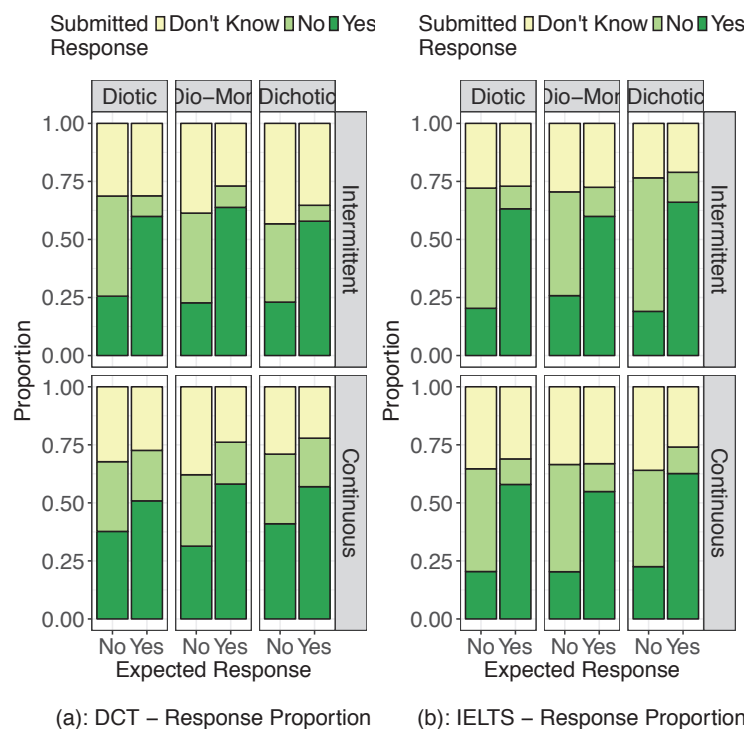


Figure 6.1 : **The Proportion of User Responses:** across the all concurrent designs.

6.3.1 Concurrent Designs Analysis

Following the protocol of performing two types of analysis, the same investigations were performed in the concurrent designs mentioned in Table 5.1. Since there were two types of contents in concurrent designs, both types of contents are discussed individually, and showed in Figure 6.1-(a) & 6.1-(b). The Figure shows that the proportion of selecting 'Don't Know' in DCT concurrent designs was higher than the baseline conditions. In DCT-Intermittent-Dichotic design, when the expected answer was 'No', the proportion of selecting Don't know was the highest, i.e. 43% and in the same design when the expected answer was 'Yes' the percentage remained 35% which is the second highest proportion. The proportion of selecting "Don't Know" when the expected answer was 'No' remained higher than the percentage when the expected answer was 'Yes'.

Similarly, as seen in the baseline condition, discussed in section 5.3.1, the per-

centage of selecting "Yes", 23%, as a wrong answer was higher than the percentage of selecting "No", 7%, as a wrong answer. The similar trends of proportion have been seen in the other designs as shown in Figure 6.1-(a). In all the designs based on DCT content, users selected all three 'Yes,' 'No' and 'Don't Know' responses as the answers to the questions.

Regarding concurrent designs based on IELTS content, the Figure 6.1-(b) shows the proportion of responses submitted by users. Similar trends appeared in the IELTS as seen in the DCT-based concurrent designs. However, the proportion of giving correct responses appeared higher and the selection of 'don't know' remained lower compared to DCT-based designs. This shows users comprehension was better in the IELTS-based designs, particularly in intermittent designs. In all the designs based on IELTS content, users selected all three 'Yes,' 'No' and 'Don't Know' responses as the answers to the questions.

The percentage of correct answers in concurrent designs were also calculated. For comparison, Figure 6.2 shows the percentage of giving correct answers in each concurrent design based on DCT and as well as IELTS content. In DCT content-based designs, the users' comprehension performance appeared low in comparison to the IELTS content-based designs. The comparison is shown in Figure 6.2 which reflects that the percentage of giving correct answers was highest in IELTS. Intermittent. Dichotic design of 63%. This percentage appeared similar ($P = 0.457$) to the baseline benchmark of 65%. The DCT. Continuous. Dichotic design appeared to be a significantly worse design ($P < 0.000$) in communicating concurrent information as the percentage of giving correct answers was as low as 39%. Among both, the continuous and the intermittent forms of concurrent designs, the intermittent form appeared better in communicating speech-based concurrent information.

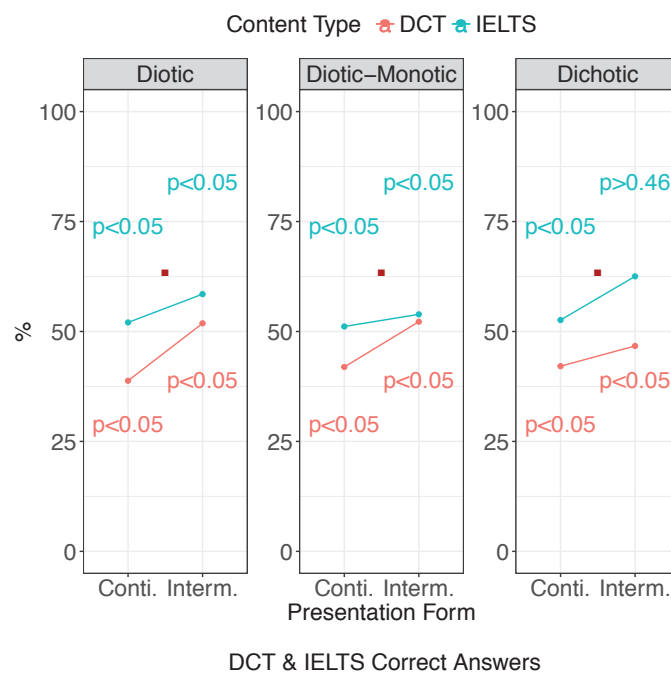


Figure 6.2 : **The percentage of correct answers:** across the concurrent designs. The *p*-values against each point show the respective statistical significance against the baseline condition.

6.3.2 Intermittent Designs in Detail

Since user's comprehension performance appeared higher in the designs based on intermittent form, this form was further investigated. In this analysis, the users' behaviour in comprehending 'Competing' questions and its comparison with the comprehension of 'Non-competing' questions in the same design was investigated. For this, those questions in the primary stream of intermittent form-based designs were marked non-competing where the relevant information content to the question was played during the silent interval in the secondary stream. All other questions were marked competing as the content related to those questions was always played in the presence of competing speech. The competing and non-competing marking is visualised in Figure 6.3.

Figure 6.4 shows the comparison between correct answer percentages in non-competing and competing questions of the DCT & IELTS content-based intermit-

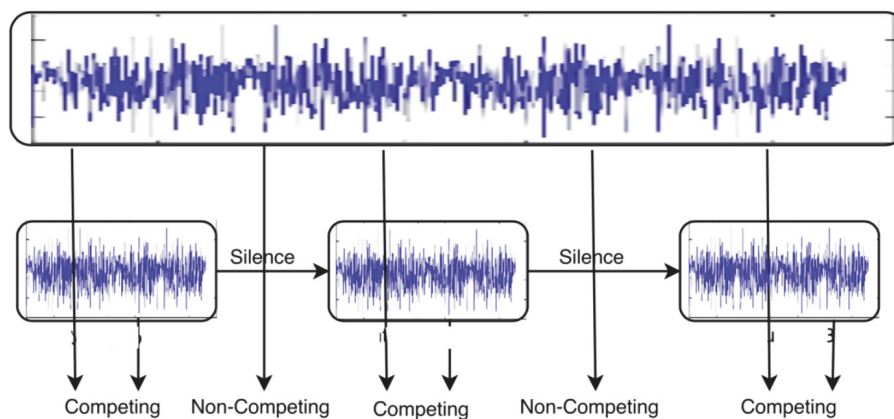


Figure 6.3 : **Competing and Non-competing Questions:** Intermittent Design indicating the 'Competing' and 'Non-competing' questions.

tent designs. The lines and corresponding p - values in Figure 6.4 indicate that the users comprehension was similar in competing and non-competing' types of questions in DCT content-based intermittent designs. In DCT-Diotic design, for non-competing questions, the percentage was 56% whereas for competing, the percentage was 51%. In DCT-Diotic-Monotic the comprehension was identical. However, in DCT-Dichotic design, the percentage of the correct answer in non-competing questions remained 40% that was lower than the percentage of correct answers for competing questions at 47%. The graph lines in Figure 6.4 related to the IELTS content-based intermittent design show that the user's comprehension was slightly higher in terms of correct answers percentage in non-competing questions comparing to the competing questions. In IELTS-Dichotic intermittent design, the correct answer percentage in non-competing question was 65% whereas in competing questions it was 62%. The same pattern of slight difference was seen in the other IELTS designs. However, these slight differences in percentage are negligible as the p - Values in Figure 6.4 show that the comprehension difference between both types of questions was statistically insignificant ($P \geq 0.05$) in all designs except in the case of IELTS-Diotic-Monotic design where p - Value was $P = 0.018$. Hence, the users' comprehension was similar in both types of

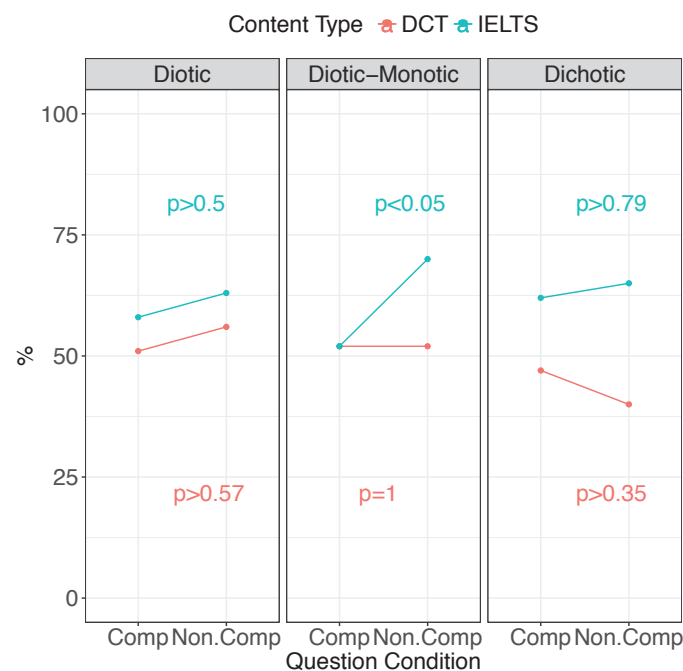


Figure 6.4 : **Percentage of Correct Answers:** with respect to competing and non-competing questions in intermittent designs.

The p-values against each point show the respective statistical significance between the correct answers against question condition.

competing and non-competing content presented in the intermittent design.

The intermittent design was a sort of combination of both, the baseline sequential communication and the continuous concurrent communication designs. In parts of the intermittent design, the portion where the silent intervals in secondary voice appeared, stimulus imitated the sequential communication whereas the other portion where the secondary stream was being played, the speech mocked the continuous communication. In other words, based on the similar information presentation, the non-competing questions shown in the Figure 6.3 were similar to the questions asked in baseline sequential communication, and the competing questions were similar to the questions asked in continuous concurrent communication.

In this analysis, first, the percentage of correct answers in non-competing questions were compared with the questions answered in the baseline condition. Figure

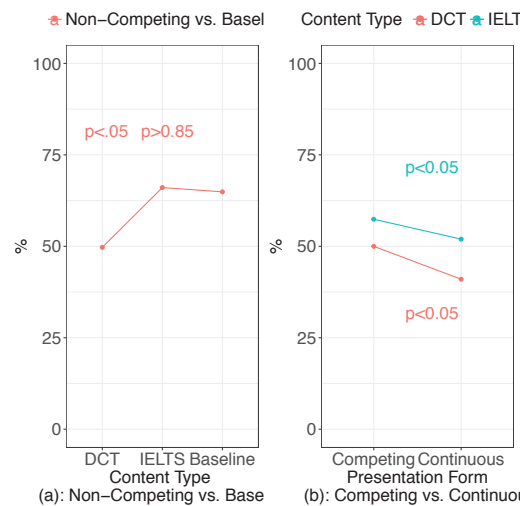


Figure 6.5 : Deep Intermittent Analysis: (a) Non-competing questions from concurrent intermittent designs compared with the baseline condition questions, (b) competing questions from concurrent intermittent designs compared with the concurrent continuous designs,

The p-values in a) show the statistical significance of comprehension in presentation forms against the baseline whereas in b) show the statistical significance of comprehension between the competing and continuous question types.

6.5 (a) shows the percentage for both content types compared with the baseline condition. The analysis showed that the percentage of non-competing questions in IELTS was almost similar, i.e. 66% to the 65% of the baseline condition. However, in DCT the percentage was significantly lower $p < .001$, 50% as compared to the baseline condition. Since the percentage of correct answers in non-competing was almost similar to the baseline condition in case of IELTS, this gave an impression that users were not overly distracted by the secondary intermittent stream in concurrent condition, particularly in IELTS content-based designs.

In the second part, the answers to competing questions in intermittent form-based designs were compared with the answers given in the continuous design of the related content. The Figure 6.5 (b) shows the results. In both DCT and IELTS, the correct answer percentage was higher in the competing questions. In DCT comparison, the percentage of giving correct answers for competing

questions was 50.31% whereas in the concurrent continuous form the percentage was 40.96%. The similar pattern appeared in IELTS content-based designs where the percentage of the correct answers for the intermittent competing questions was 57.4% whereas in the concurrent continuous condition it was 51.96%. The p – value shown in Figure 6.5 indicates the significance difference of giving correct answers between competing and continuous questions. Since $P < 0.05$ in both the types, therefore, comprehension was significantly better in competing questions asked in intermittent concurrent design compared to the continuous concurrent design.

6.4 Discussion

Figure 6.2 reflects that the user's comprehension was better in intermittent form compared to the continuous form of delivery. In all six designs, users percentage of correct answers was higher in intermittent form compared to the continuous form. It indicates that the intermittent approach provided ample time to the users to understand the context and details of the continuous information that created bandwidth for the user to listen to the intermittent speech by compromising the continuous speech. This assumption leaves a question, how much information have users obtained and how much they had compromised from both the information streams. An interesting analysis can be carried out to investigate this aspect from the same result dataset.

Besides the intermittent form of presentation, an increased comprehension behaviour was witnessed when the spatial differences were involved additionally in speech-based streams. The designs, particularly those based on IELTS-content, showed that spatial difference played a significant role in comprehending the information. The comparison showed in Figure 6.1 reflects that the percentage of correct answer was highest. According to the results, the better comprehension comparable to baseline condition can be achieved by providing concurrent

information intermittently and in the Dichotic condition.

In Figure 6.1-(a), illustrating the proportion of responses for all designs, the percentage of selecting 'Yes' as a wrong answer was higher than the percentage of selecting 'No'. Users inclination towards selecting 'Yes' for the higher number of times was based on a user's instinct towards agreeing with questions when they 'didn't know' the answers. It implies that the absence of 'Don't Know' option could have led to less accurate comprehension calculation.

The results in the designs based on DCT content were inconsistent compared to the IELTS content-based designs because of the number of factors that include:

- The audio quality of the mono channelled DCT was not as clean in listening as was in stereo channelled IELTS.
- The content was natively played in low-pitched (male) voice that was converted into the high-pitched (female) voice for six files to attain the discrimination on the basis of gender (fundamental frequency) voice.
- The continuous stories were broken into the chunks to answer the pre-set questions designed natively.

These factors broke the continuity of the discourse/story and audio quality. In the case of IELTS, these challenges were not faced as a sufficient number of files were available in both the male and the female files, and the audio quality was stereo. Besides this, the questions for IELTS were custom created. Therefore, the broken continuity of the discourse/story challenge did not appear.

6.5 Limitations & Future Work

This analysis shows the potential of communicating concurrent information with suitable information design but does not fully cover all the aspects in concurrent condition. This analysis does not adequately cover user comprehension behaviour.

For example, at what point users switched their attention to the secondary stream and how long the attention persisted. During switching the attention, how much information have users lost from the primary information stream while focusing on the secondary stream and vice-versa? In chapter 7, the analysis is extended to the next level and evaluated comprehension depth of both the primary and the secondary information streams in speech-based concurrent information designs.

An earlier version of the research discussed in this chapter has been presented in the following paper:

Publication 5: M. A. u. Fazal, S. Ferguson, and A. Johnston, “Investigating Concurrent Speech-based Designs for Information Communication,” in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ACM. New York, NY, USA: ACM, 2018, pp. 1–8.

Attached as Appendix-K.

Publication 6: —, “Investigating Concurrent Speech-based Designs for Efficient Information Communication - Extended Analysis,” *Journal of the Audio Engineering Society (JAES)*, vol. -, no. -, pp. 1–8, 2019, submitted.

Attached as Appendix-L.

Chapter 7

Evaluation of Information Comprehension Depth in Speech-based Concurrent Designs

This chapter further extends the analysis of the experiment discussed in chapter 5, and reports another aspect of interest regarding speech-based concurrent information communication. This analysis evaluates the comprehension depth of information by comparing comprehension performance across several different formats of the questions (main/detailed, implied/stated) in each concurrent speech-based design. The analysis determines whether users, in addition to being able to answer the main questions, are successful in answering the implied questions, as well as the questions that required detailed information.

7.1 Aims & Motivation

Aims: The third aim of the study is to evaluate comprehension depth of both the primary and the secondary information streams in speech-based concurrent information designs and determines which design was the most effective in communicating speech-based information concurrently. Additionally, the analysis is performed to satisfy the following questions: a) Does the pattern of comprehension depth remain similar to that seen in a baseline condition, where only one speech source was presented? b) Does the Diotic-Monotic concurrent design exploiting REA (right ear advantage) provide an advantage in content comprehension? c) Does the audio quality impact the comprehension? d) What do users report about their experience in concurrent information communication? e) Lastly, do male and female users show equal interest in concurrent information

communication?

Motivation: The motivation for this analysis is to produce speech-based concurrent information designs for auditory displays that could efficiently communicate concurrent information similar to the performance that people achieve in conventional sequential information communication. Such multi-channelled information can help to contribute to understandings of voice-based interaction methods and leverage new forms of human-computer interaction in a computing environment.

7.2 Method

Discussed in section 5.2.

7.3 Results

An analysis was carried out that started by comparing the overall comprehension of content in the primary stream with the overall comprehension of content in secondary streams and then extended to determine the depth of comprehension in each stream of speech-based concurrent information designs. The comprehension depth is defined as understanding spoken information from the scale of main to the detailed levels. The arrangement of questions in MIS, MII, DTS, and DTI categories and users' correct answers to these categorised questions determined the comprehension depth in both the primary and the secondary streams in each stimulus design. The analysis regarding comprehension depth carried out by comparing each concurrent design with the baseline condition. The analysis also covered the qualitative response submitted by the users explaining their experience, emphasising the cognitive load that was incurred when interacting with the speech-based concurrent information designs.

The results of overall comprehension comparison between primary and secondary streams for each design followed by the comprehension depth results for

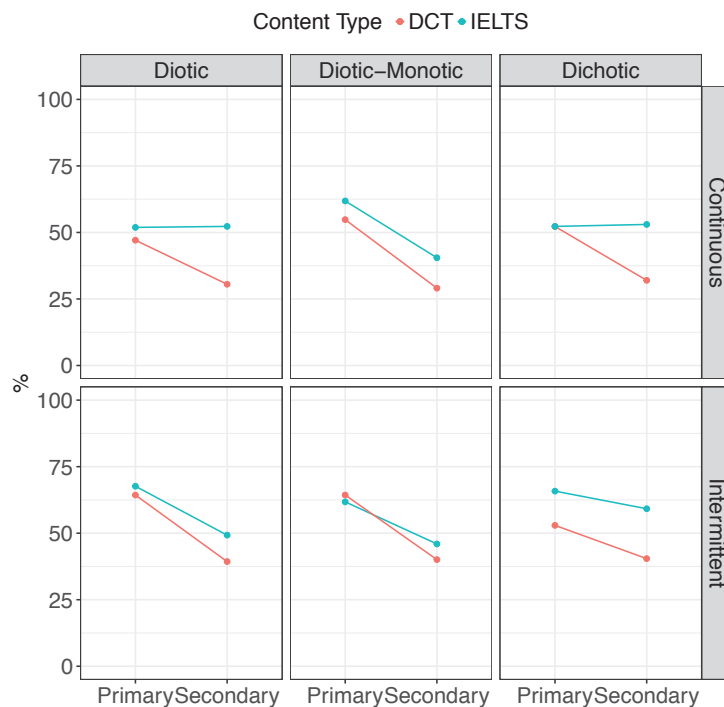


Figure 7.1 : **Percentage of Correct Answers:** with respect to the primary and the secondary streams in concurrent speech-based designs.

concurrent designs are discussed in sub-sections 7.3.1 & 7.3.2. The qualitative response submitted by users is discussed in sub-section 7.3.3.

7.3.1 Overall Comprehension Comparison between Streams

The first part of this analysis discusses the comprehension comparison between the primary and the secondary streams for each design. To reflect the comparison, the percentage of correct answers for each concurrent speech-based design of each stream is shown in Figure 7.1. The users' response to a question that matched to the expected answer counted as a correct answer, whereas if the answer was not the expected answer or the 'Don't Know' option was selected, these were considered as a wrong answer. In DCT content type, as shown in Figure 7.1 and *p-values* mentioned in Table 7.1, the comprehension was significantly higher in the primary stream in comparison to the secondary stream. The comprehension

difference was unchanged in both the Continuous and the Intermittent forms of the concurrent speech-based designs.

Table 7.1 : **Proportion Comparison Test:** of correct answers between the primary and the secondary streams, with Bonferroni correction, * * * < 0.001

Design	p-Value
DCT	
DCT Continuous Diotic	***
DCT Continuous Dio-Mon	***
DCT Continuous Dichotic	***
DCT Intermittent Diotic	***
DCT Intermittent Dio-Mon	***
DCT Intermittent Dichotic	0.005
IELTS	
IELTS Continuous Diotic	1
IELTS Continuous Dio-Mon	***
IELTS Continuous Dichotic	0.932
IELTS Intermittent Diotic	***
IELTS Intermittent Dio-Mon	***
IELTS Intermittent Dichotic	0.132

On account of all the DCT content-based concurrent designs, the average percentage for the primary stream was 56% whereas in the secondary stream it was 35%. In designs based on IELTS content, in three designs - Diotic Continuous, Dichotic Continuous and Dichotic Intermittent - the comprehension was the same, and in rest of the three designs the comprehension was higher in the primary stream than the secondary stream as shown in Figure 7.1 and *p-values* mentioned in Table 7.1. On account of all the IELTS content-based concurrent designs, the average percentage for the primary stream was 60% whereas in the secondary stream it was 50%. In conclusion, users comprehension mostly was higher in the primary streams compared to the secondary streams in the speech-based concurrent designs.

7.3.2 Comprehension Depth for Concurrent Condition

In this section, the evaluation of the comprehension depth of the content in concurrent designs is evaluated. Figure 7.2 shows the percentage of correct answers in both the streams for MIS, MII, DTS, DTI for each design individually. Each MIS, MII, DTS, DTI data point (percentage) of each concurrent speech-based design is statistically compared with the relevant data point in the benchmark set from the baseline condition and are mentioned in Table 7.2. In almost all of the speech-based concurrent designs, the comprehension was significantly lower than the benchmark except in one design IELTS.Intermittent.Dichotic. In the primary stream of IELTS.Intermittent.Dichotic design, the percentage of correct answers to the questions set from MIS was 85% whereas in MII, DTS, DTI it was 68%, 56%, 54% respectively. In the same IELTS.Intermittent.Dichotic design, the secondary stream's percentage for MIS was 76% whereas in MII, DTS, DTI it was 57%, 53%, 50% respectively. The results show that the user's comprehension depth in this concurrent speech-based design was similar to the comprehension depth calculated in the benchmark.

This section briefly discusses the user's experience of interacting with the concurrent speech-based design. In addition to questions based on content, a descriptive question, "Can you please share your experience in using this system?" was asked in order to gain an understanding of the user's experience with the system. The users' reactions and suggestions are concisely presented to discuss the viability of concurrent speech-based communication. In turn, this provides several hints to explore the possibility of communicating speech-based information concurrently.

A few users reported that their experience was fairly interesting and intriguing as the method carries the potential to improve the multitasking perspective of life, subject to better implementation and considerations of design. The concurrent approach can be useful in contexts where information is not critical – for instance,

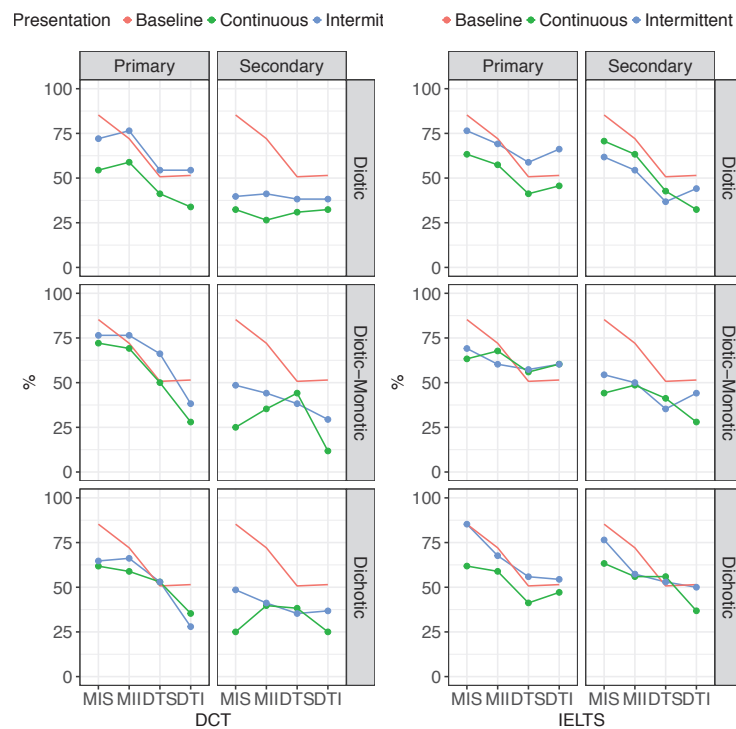


Figure 7.2 : **Users' comprehension:** in each stream with respect to MIS, MII, DTS, DTI.

listening to stock market reports, sports commentary and news reports. Another example could be listening to music or sounds, rather than densely layered narratives.

Users also reported that they felt that the use of these systems depends on an individual's mental capabilities and differing preferences. This infers that controls to configure the precise format for how to listen to concurrent information should be given to the users so that they may set up the listening session according to their preferences. Users pointed out that the female voice felt more clear and dominant as opposed to the male voice, which can be partly a result of the employed reproduction equipment frequency response. The secondary voice presented intermittently was less challenging cognitively and was easier to comprehend, however, it created an issue of excessive attention switching. Moreover, it was reported by users that the content played with higher play-rate

Table 7.2: **One-to-One Proportion Comparison Test Results:** between the correct answers in each stream of the baseline and the concurrent designs with respect to MIS, MII, DTS, DTI, with Bonferroni correction, * * * < 0.001

Concurrent Design	Primary Stream				Secondary Stream			
	MIS	MII	DTS	DTI	MIS	MII	DTS	DTI
DCT								
DCT Conti. Diotic	***	0.082	0.256	0.026	***	***	0.011	0.015
DCT Conti. Dio-Mon	0.038	0.79	1	0.002	***	***	0.46	***
DCT Conti. Dichotic	***	0.082	0.879	0.042	***	***	0.125	0.001
DCT Intermi. Diotic	0.038	0.614	0.729	0.804	***	***	0.124	0.102
DCT Intermi. Dio-Mon	0.173	0.614	0.052	0.102	***	***	0.124	0.004
DCT Intermi. Dichotic	0.001	0.482	0.882	0.002	***	***	0.053	0.066
IELTS								
IELTS Conti. Diotic	0.001	0.052	0.257	0.525	0.022	0.265	0.35	0.015
IELTS Conti. Dio-Mon	0.001	0.634	0.579	0.292	***	0.002	0.257	0.002
IELTS Conti. Dichotic	***	0.083	0.257	0.662	0.001	0.032	0.579	0.067
IELTS Intermi. Diotic	0.173	0.785	0.346	0.065	***	0.019	0.082	0.4
IELTS Intermi. Dio-Mon	0.011	0.123	0.457	0.297	***	0.003	0.053	0.4
IELTS Intermi. Dichotic	1	0.625	0.586	0.804	0.173	0.051	0.882	0.961

felt a better approach to communicating speech-based content fast as it did not affect the ability to focus on stream and remember the content. User observations regarding female voice clarity and intermittent communication proved evident in the computational analysis as the comprehension recorded was better in both the cases.

7.3.3 Users' Experience

Some users completely disagreed with the idea of communicating speech-based information concurrently. They reported that the communication in parallel might mean that users miss significant information. Users also reported that it was difficult to focus on the concurrent streams at the same time. Distractions made them confused and resulted in overlapping of the content from both the streams. Moreover, the identical words and phrases played together in concurrent streams made it confusing to comprehend information. Additionally, some users reported the experiment was extremely challenging, especially when the task required to

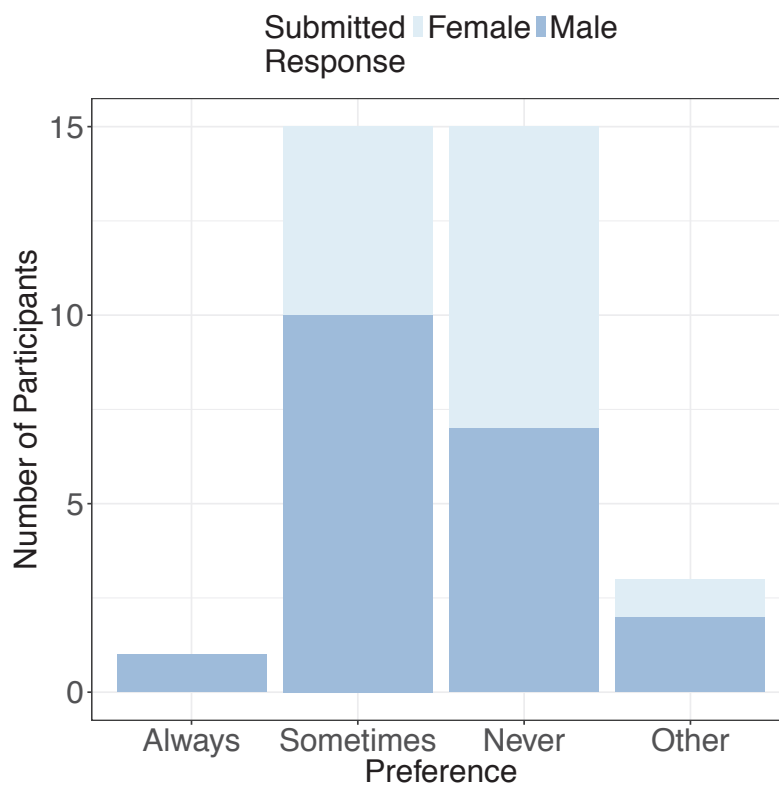


Figure 7.3 : **Participants Preference:** for concurrent information communication

not only listen to the streams but also answer questions.

Finally, based on their experience in the experiment, participants were asked how often they would prefer concurrent communication over the baseline sequential information. Half of the users as shown in Figure 7.3 opted 'sometimes' whereas others said 'Never' and one said 'Always' which suggest cognitive load was an issue that needs investigation. Regarding the selection of 'Sometimes' to this question, there is a need to identify the contexts where users would be looking to preference concurrent communication over sequential.

7.4 Discussion

Compared to the baseline condition, user's comprehension was significantly lower when both streams were provided continuously in speech-based concurrent design.

In all the conditions where the concurrent information streams were presented continuously, the comprehension of MIS was significantly lower than the comprehension of MIS in the baseline condition. However, comprehension improved when additional design factors were involved in concurrent speech-based designs. For instance, the secondary information stream provided intermittently in speech-based concurrent design having stereo-channelled audio quality rendered better concurrent-speech comprehension compared to the continuous presentation.

In addition to the above factors, when the information was further facilitated with a dichotic spatial cue, it attained similar comprehensibility that was achieved in the baseline condition. This can be seen in results (section 7.3.2), the comprehension was higher in IELTS.Intermittent.Dichotic Design. In this design, the information comprehension was similar not only in MIS but also in MII, DTS, and DTI. In conclusion, the concurrent speech-based information works better if the information is provided intermittently, dichotically and also if the audio quality of the information stream is good (stereo-channelled).

In all intermittent designs and those where information was presented by involving Diotic-Monotic spatial difference, user's comprehension was better in the primary stream than the secondary stream. In these designs, users considered the female voice (higher fundamental frequency) as a primary voice to focus because of the following reasons:

Continuity The primary stream in female voice was continuous whereas the male voice was played intermittently. The continuity of stream affected the user behaviour to treat the female voice as a primary voice.

Sound Pressure Level In Diotic-Monotic designs, the female stream was dominant as it was coming to both ears comparing to the male stream that was coming to right ear only. The difference in the loudness contributed to treating female voice as a primary voice to pay attention.

These findings may help to communicate two streams concurrently, one treated as a primary voice attracting more attention of the users and the other to be treated as secondary information to provide additional information.

In this study, besides diotic and dichotic, another spatial cue combination, Diotic-Monotic, was introduced where the primary stream was played in both ears and the secondary stream, incited from the Right Ear Advantage (REA), was played in the right ear only. Aided by REA, it was expected that the secondary stream information would require less attention or processing for comprehension and users would be able to pay dominant attention to the primary stream played in both the ears. Consequently, this design would render better comprehensibility.

The results showed that this approach does not produce an advantage. In fact, in one of the designs, DCT.Continuous.Diotic-Monotic, the comprehension of secondary stream was the lowest. The fundamental reason is the low intensity of the secondary stream comparing to the primary stream. Since the secondary stream in this design was coming to one ear only, therefore, the loudness of this stream was perceived lesser than the primary stream coming to both ears. However, when loudness was the same for both streams in Dichotic designs, the comprehension of secondary stream remained better than the Diotic and Diotic-Monotic designs. In Dichotic, the same loudness was not the only factor; the other important cue was that both streams had the spatial difference of 180 degrees which also contributed to attaining better comprehension. Considering this, an interesting investigation to explore REA could be to increase the loudness of the secondary stream presented to right ear only and bring it equal to the loudness of the primary stream presented to both ears and then examined. In the present experiment, the Diotic-Monotic design did not serve any advantage in speech-based concurrent communication.

As mentioned in the methods section, in DCT content the default lower fundamental frequency (male) voice was increased by 17% to generate the im-

pression that the other stream is being played in a female voice. Considering the DCT.Continuous.Diotic design, where the only difference between both the streams was a difference of values in fundamental frequency, users comprehended more information from the content played in higher fundamental frequency. The result shows that the high-frequency voice attracts more attention of the listeners in case of competing voices compared to the low-frequency voice. The application of this finding could be to use high-frequency voice in a complex sound environment to disseminate the critical information that requires the immediate attention of the listeners among the competing voice-based streams.

Moreover, as discussed in the methods section, the questions were arranged in 4 categories MIS, MII, DTS, DTI and formed on the basis of information repetition in the content to assess the comprehension depth. It was expected that the user's content comprehension would remain in the same order as mentioned above. The main information was repeated multiple times within the content, whereas the detailed information was played only once in the content. User's comprehension was high in MIS and MII and low in DTS and DTI. The analysis showed that users comprehended the main information well and were able to comprehend information about 50% in the baseline condition. In almost all the concurrent speech-based designs the percentage of correct answers was significantly lower than the baseline condition. However, the pattern of comprehension depth was similar in both streams played concurrently in each speech-based concurrent design as was seen in the baseline condition. Users answered more questions correctly which were drawn from MIS/MII and performance was lower in DTS and DTI. In conclusion, the pattern of comprehending information did not change in concurrent speech-based design.

Regarding information presentation preference, 15 users said they would 'sometimes' prefer concurrent speech-based communication over sequential, and one said they would prefer it 'always'. 10 males and 5 females expressed an interest.

Male users showed a higher interest in speech-based concurrent communication than female users.

7.5 Limitations and Future Work

Many users reported high cognitive load in concurrent speech-based information communication. Since the objective of this study was to assess the content comprehension by the users in concurrent speech-based communication, this experiment required users to listen content from 14 stimuli designs and answer the questions from the content. The study impacted high cognitive load and demanded extensive use of memory. The chapter 8 focuses on user experience in concurrent communication, and subjectively evaluates various combinations of information streams in concurrent speech-based information communication using an experiment to identify the best-suited combinations of information types for concurrent communication.

An earlier version of the research discussed in this chapter has been presented in the following paper:

Publication 7: M. A. u. Fazal, S. Ferguson, and A. Johnston, “Evaluation of Information Comprehension in Speech-based Designs for Concurrent Audio Streams,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. -, no. -, pp. 1–18, 2018, submitted.

Attached as Appendix-M.

Chapter 8

Evaluating Various Combinations of Information Streams in Concurrent Information Communication

This chapter reports on another experiment that comprehensively investigates the cognitive workload experienced when listening to a variety of combinations of information types in concurrent formats. Fifteen different combinations of information streams were investigated, and the subjective listening workload using the NASA Task Load Index (TLX) (Hart and Stavenland, 1988; NASA, 2018a) was calculated for each of these combinations. This approach allows us to determine which types of information are best suited, in terms of cognitive workload, for concurrent communication. It determines how much users preferred each combination and how frequently they would use these combinations. It is expected that this experiment will help digital content creators and designers to communicate information to users more efficiently. In this study, besides the speech-based information streams, songs and non-vocal music (instrumental) were included. The addition is to determine the impact and user experience while listening to a speech-based information stream combined with music or a song stream dichotically, as may well be experienced in an ecological setting.

8.1 Aims & Motivation

Aims: This experiment aims to comprehensively analyse the cognitive load by subjectively calculating workload that users endure while listening to two different audio streams concurrently. The analysis is performed to satisfy the following questions: : a) Does the cognitive workload remain similar in each type of combi-

nation? b) Do users respond differently regarding different combinations in terms of preference and frequent use? c) Which information type(s) is preferred the most by users in concurrent combinations? d) Does the intermittent form of communication create a lower cognitive workload index in speech-based information communication, when compared to the continuous concurrent form?

Motivation: The motivation behind conducting this experiment is to determine the viability of communicating concurrent information by experimenting with various combination of streams. Since the study involves dichotic presentation of songs and non-vocal music with other information types not having background music, it is therefore expected that the study will not only provide a path to deliver information quickly but would also enhance the user experience when listening to speech-based information, for example, a documentary, interview or commentary combined with randomly associated / user's selected music or song. Moreover, the intermittent design in terms of evaluating cognitive load is also being tested, and it is therefore expected that this method may help in designing the human-computer interaction by mapping it to overlay type of techniques of GUI as explained in the introductory Chapter 1.

8.2 Method

The method adopted for this experiment is outlined below.

8.2.1 Participants

After receiving institutional Human Research Ethics Committee approval for the research protocol, attached as Appendix-E, user participation campaigns were launched. The participants were selected based on two criteria: 1) not having a significant hearing impairment, and 2) having competent English language skills, as the listening experiment's content was in the English language. The users,

selected for participation, were offered gift cards worth 30 AU\$ each. In total, 40 participants, 20 female, and 20 male took part in the experiment after providing consent. The mean age of the participants was 23 with the standard deviation of 6.

Table 8.1 : **Combinations of Different Types:** of Information Streams in the Concurrent Stimuli

Streams in the Right Ear	Monolog	Interview	Comment.	News	Songs	Music
	Streams in the Left Ear					
Monolog	-	-	-	-	-	-
Interview	✓	-	-	-	-	-
Commentary	✓	✓	-	-	-	-
News Headlines	✓	✓	- ✓	-	-	-
Songs	✓	✓	✓	✓	-	-
Music	✓	✓	✓	✓	✓	-

8.2.2 Design

Concurrent Condition

In this condition, two different types of information streams were concurrently communicated in a dichotic form to users. For concurrent communication, a series of stimuli were created where each stimulus was created by combining the two different types of information streams. One stream was presented in the right ear, and the other stream was presented in the left ear. This dichotic presentation (spatial separation) was achieved using the panning feature in an open source software Audacity (Audacity).

For concurrent stimuli, combinations were created from the types of information streams that included: Monolog (Documentary), Interview (Dialog), Commentary (Football), News Headlines, Song (Vocal), Music (Non-Vocal). Each type of information stream was combined with the other types of information streams once. For simplifying the explanation, each of the rendered concurrent stimuli design is described in Table 8.1 and illustrated in Figure 8.1 for further clarity.

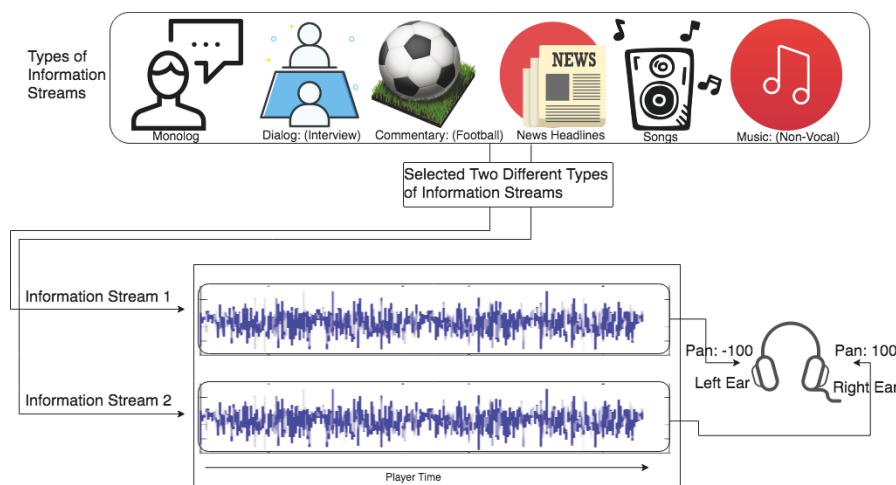


Figure 8.1 : Concurrent Stimulus Design

Intermittent In all the stimuli, the information streams were presented continuously, except for the stimuli where an information stream was combined with the news headlines information stream. Following the form illustrated in Figure 5.1 of the previous study discussed in Chapter 5, the news headlines stream was manipulated and transformed to intermittent from the continuous information presentation. For this, the news headlines bulletin was broken into temporal segments, and after each news headline, a gap of 15 seconds of silence was added. This intermittent form was involved because in the previous studies it was found that the users' comprehension was the best in concurrent information communication formats. Therefore, the information streams combined with the news headlines information stream was of an intermittent concurrent design type. The rest of the stimuli were based on continuous information design.

Baseline Condition

In this condition, no concurrency was involved. A type of information stream was randomly selected and sequentially communicated to the users in a conventional form. The purpose of this condition was to set a users' experience benchmark, and later use it to compare the experience with concurrent conditions.

8.2.3 Material

First, the streams of six information types were selected that included: Monolog (documentary), Dialog (interview), Commentary (football), News Headlines, Songs (vocal) and Music (instrumental / non-vocal). For each information type, BBC channels on YouTube were searched to find quality information presentations. For each information type, six videos of a maximum of 2 minutes duration were selected, except for the commentary information type. For commentary, the first twelve minutes of the sports match was broken in 6 equal (in duration) files. Following the selection method, there were 36 files in total.

Based on the ecological choices, the *monolog* streams were selected from the BBC program *Lip Service* wherein each selected documentary a woman discussed a trait of her life. Interviews (*dialog*) were selected from the BBC's program *BBC Celebrity Interview* where a male host interviewed a male celebrity. The sports *commentary* was from the BBC 5 Live Radio and was recorded in the male voice covering a football match between Napoli and Manchester City. The *news headlines* spanned six different dates and were selected from the BBC World News. Three news headlines were recorded in the female voice and three in the male voice. The *songs* were selected from the BBC Radio-1 Channel where three of the singers were female and three male. For the *music*, the background music of the Hollywood movie *Viceroy* composed by the Academy Award winner AR Rahman was selected.

8.2.4 Stimuli Information

Since each type of information stream was combined with the rest of the types of information streams, users were presented with 15 different concurrent combinations. Besides the concurrent combinations, a baseline stimulus was also presented. Hence, a user was presented with the 16 different stimuli (15 Concurrent, 1 Sequential).

In total, there were 576 stimuli combinations that were created, including the

stimuli representing the baseline condition in order to remove the combinational effect. A user was presented with 15 stimuli, each representing each combination, as well as 1 that was the baseline stimulus. In this randomisation, the combinational effect was removed to make sure users were not provided information streams that repeated information types. The order of presenting combinations was random to remove the ordering effect.

8.2.5 Measures

The duration of each stimulus was 2 minutes. After listening to each stimulus, users were presented with a questionnaire, attached as Appendix-F to share their experience. The experience was obtained using the NASA-TLX subjective, multidimensional assessment tool (Hart and Stavenland, 1988; NASA, 2018a). NASA-TLX is a standardised tool that rates perceived workload in order to assess a task system, or a team's effectiveness or other aspects of performance. Besides being cited in over 4400 research studies, this tool has been used in many research studies investigating concurrent communication (Parente, 2008; Vazquez-Alvarez et al., 2015; Hinde, 2016; Truschin et al., 2014; Vazquez Alvarez and Brewster, 2010; Towers, 2016; Vazquez-Alvarez et al., 2014). The test has two parts. In the first part, the total workload is measured using the following NASA (2018b) subjective subscales:

1. "Mental Demand - How mentally demanding was the task?"
2. Physical Demand - How physically demanding was the task?"
3. Temporal Demand - How hurried or rushed was the pace of the task?"
4. Overall Performance - How successful were you in accomplishing what you were asked to do?"
5. Effort - How hard did you have to work to accomplish your level of performance?"

6. Frustration Level - How insecure, discouraged, irritated, stressed, and annoyed were you?"

In order to gain more information about the user experience, two questions were added:

7. Like (Preference) - How much did you like this combination ?
8. Frequent - How frequently will you be using this combination of streams?

The second part of NASA-TLX intends to create an individual weighting of the above mentioned six subscales by letting the subjects compare them pairwise, based on their perceived importance. After this, some arithmetic operations are used to be performed in order to compute the perceived workload index.

8.2.6 Apparatus

In order to minimise participation time, a web-based system using PHP, MySQL, JQuery, HTML5, CSS, and Bootstrap was developed to play the stimuli. The web system was accessible via any latest web browser. 16 HTML audio players were designed to play each stimulus design that was presented in steps. Steps refer to the 2nd audio player that was shown to the users when they first completed the hearing of the first audio player and submitted their experiential response using NASA-TLX form. The NASA-TLX forms were created using HTML, and users submitted their response online directly into a database.

The tests were conducted in a quiet purpose-built room in the Creativity and Cognition Studios (CCS) of the University of Technology, Sydney. Three identical Apple iMac computers, having 2.7GHz quad-core Intel Core i5 processor, 8GB RAM, installed with Yosemite 10.10.5 OS were arranged in the studio. To listen to the audio stimuli Beyerdynamic's DT770 250 OHM headphones were used that were connected to the headphone jack of the computer. Users were provided with the gain control only to set the intensity as per listening choice. Since the three

computers were used in the studio, up to three participants were engaged in the study simultaneously.

8.2.7 General Procedure

The selected users were verbally briefed on the study protocol before the start of the study. Instructions were presented on a screen after registration. Before starting the study, users entered their demographic profile information that included, name, email, age, gender, primary language, qualification, profession, country, mood and hearing/visual impairment status. At the end of the study, users' detailed responses relating to their experience of information communication. Users' responses were stored in a MySQL database for post-experiment analysis.

The entire experiment interface including form obtaining user's demographic profile information, playable URLs of stimuli representing each combination, and NASA-TLX based questionnaire is produced in Appendix-F.

8.3 Results

There were four aspects to the analysis of results. First, the baseline condition was compared to the concurrent communication. Second, the analysis was extended by determining the subjective workload index for each combination, as compared to the baseline condition. Third, the impact of each information stream type on the user's experience when combined with the rest of the information stream types was explored. Fourth, the workload index and other experiential observations of the information stream types with respect to their presentation in the left ear and the right ear were determined. All four analysis steps are discussed in the subsequent subsections.

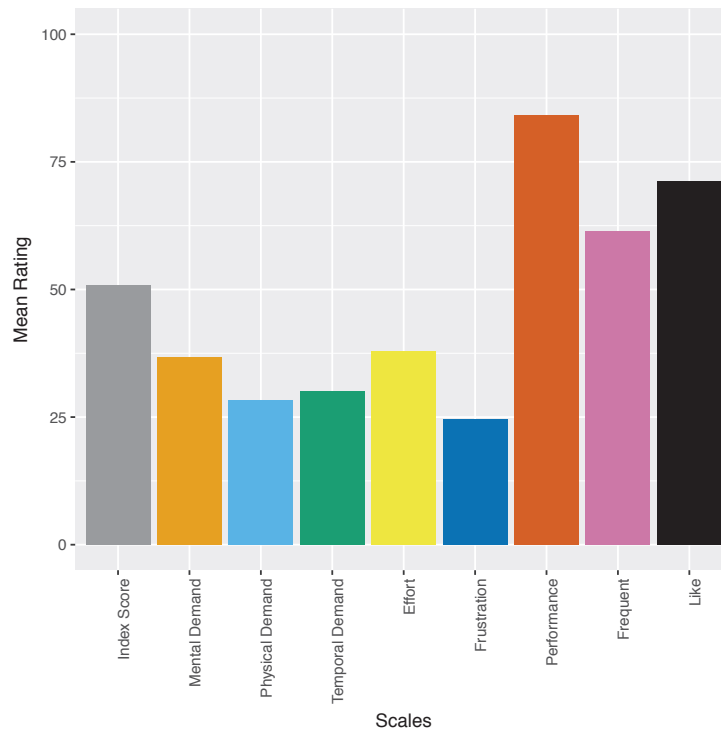


Figure 8.2 : **User Experience in Baseline Condition:** shows perceived workload index score for listening task and ratings for subscales, including the frequent and preference (like) scales.

8.3.1 Baseline vs. Concurrent

We commenced an analysis on the baseline condition and calculated the mean ratings for each subscale, along with determining the NASA-TLX workload index. After this, the same procedure was performed with the concurrent communication as a whole, without taking the combination types into account and, finally, a comparison was completed between the concurrent and the baseline condition.

Baseline Condition The baseline mean scores for each rating scale is shown in Figure 8.2. The mean rating for the mental demand in baseline condition was 36.75, whereas, for physical demand, temporal demand, effort, frustration and performance was 28.25, 30.00, 38.00, 24.50, 84.12 respectively. Using these subjective subscale ratings, combined with the weighting measure of the NASA-TLX, the calculated mean index score for baseline listening task appeared 50.81.

Similarly, regarding the frequent scale, that is asking users how frequently they would listen to the baseline condition, the mean rating was 61.37. For the baseline condition, the mean preference rating was 71.12%. These ratings set a benchmark that was used to draw comparisons with the concurrent combinations.

Concurrent Communication Following the pattern adopted in the baseline condition, the mean score was calculated for each scale, as illustrated in Figure 8.3. As shown in the Figure 8.3, the mean rating for the mental demand in baseline condition was 55.24, whereas for physical demand, temporal demand, effort, frustration, and performance was 40.87, 46.27, 56.72, 42.92, 62.82 respectively. Using these subjective subscale ratings combined with the weighting measure of the NASA-TLX the calculated mean index score for concurrent listening task appeared 59.12. Similarly, asking users how frequently they would listen to the concurrent condition, the mean score was 37.06, and the mean score for their preference of the concurrent condition was 42.07.

To statistically compare the mean concurrent ratings with the baseline condition, two-way *analysis of variance* (ANOVA) test (Copenhaver and Holland, 1988) was used. The ANOVA results, mentioned in table 8.2, showed that the presentation type (baseline — concurrent condition) does not have a significant impact on user response, $F(1, 5742) = 2.359, p < 0.125$. However, the interaction between the presentation and rating scales, $F(8, 5742) = 27.098, p < 0.01$, had a significant impact on user response.

Since the interaction between the presentation and rating scales was significant, the *Post hoc* Tukey HSD analysis (Miller, 198; Yandell, 1997) was performed to compare each concurrent mean rating with the baseline mean rating. The results showed that the index score regarding the baseline condition and the concurrent condition does not have a significant difference. However, all rating scales, except physical, appeared significantly different in the baseline condition and the

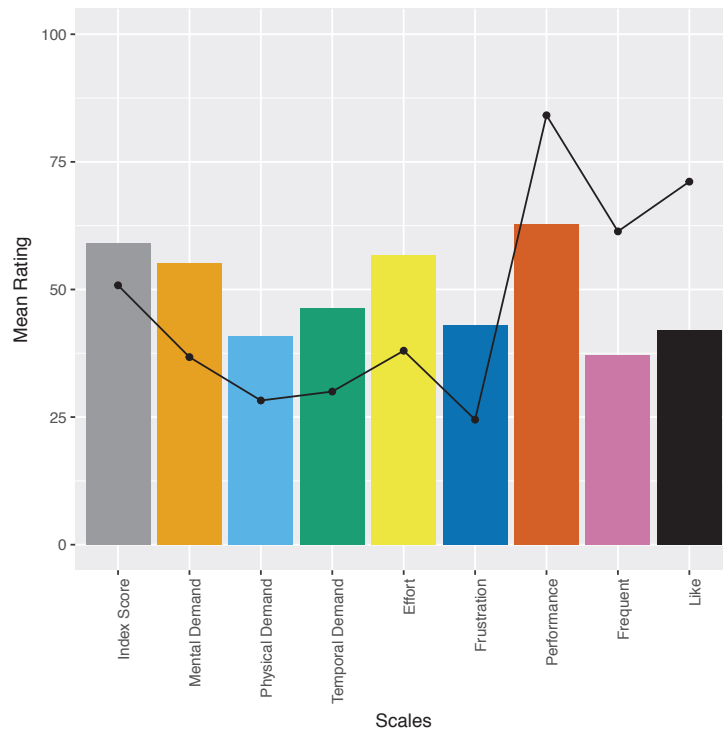


Figure 8.3 : **Users' Experience in Concurrent Communication:** compared with the baseline condition shown with the black continuous line.

concurrent condition ($p < .05$). Users' responses showed that they preferred the baseline condition, and, therefore, would more frequently use it when compared to the concurrent condition. Table 8.2 shows the statistical difference between the concurrent condition and the baseline condition for each rating scale.

Table 8.2 : **Post hoc Tukey HSD Analysis:** ($p - values$) comparing mean ratings of concurrent scales with the relevant baseline scales: (*Signif.codes* : $\leq 0.001 = ***$, $\leq 0.01 = **$, $\leq 0.05 = *$)

Index	Mental	Physical	Temporal	Effort	Frustration	Performance	Frequent	Like
0.81	***	0.12	**	***	***	***	***	***

8.3.2 Concurrent Combinations

The results of each combination type have been compared with the baseline condition. The mean scores of the scales for all the combination designs are

individually illustrated in figure 8.4. The comparison of each of the mean scores with the baseline condition is also depicted using a continuous black line indicating the mean values in the baseline condition.

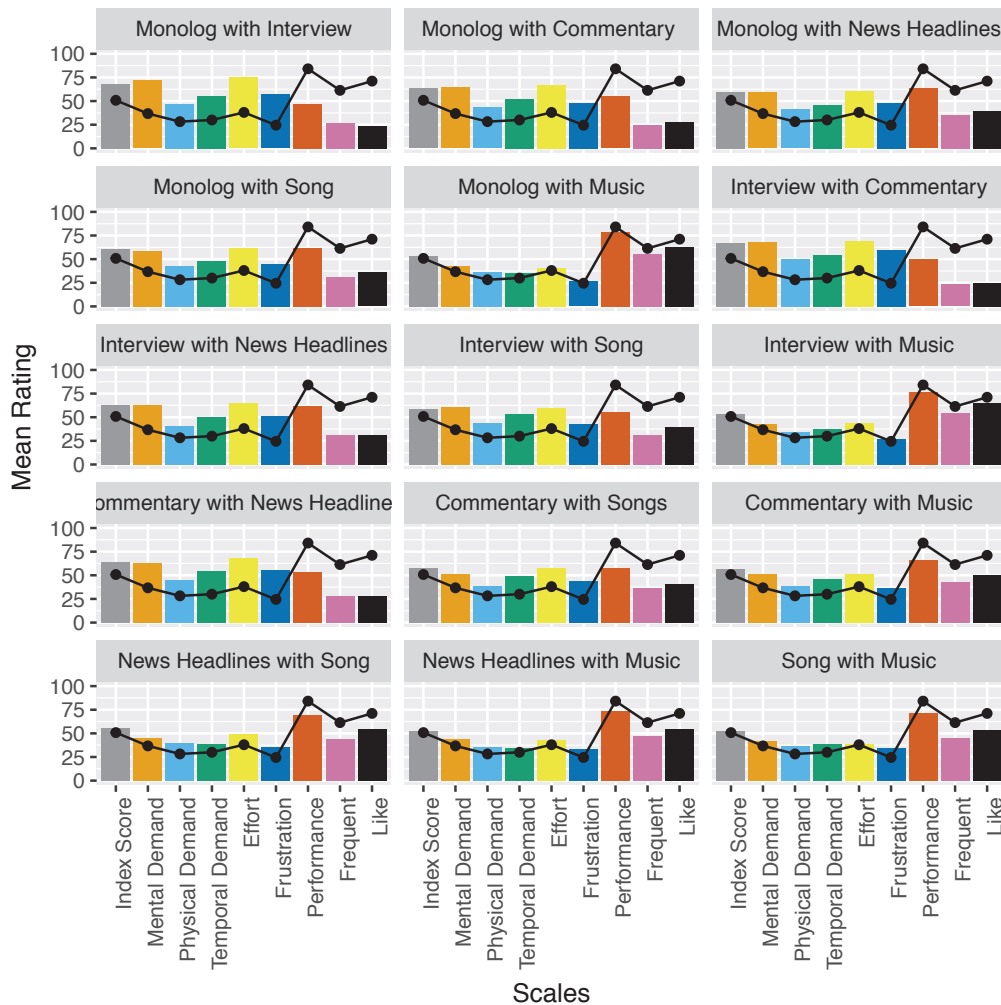


Figure 8.4 : **Users' Experience in each Combination:** of concurrent communication, also compared with the baseline condition shown with a continuous line.

Besides illustrating the results in Figure 8.4, the concurrent combinations with the ANOVA results are discussed in the following subsections.

To discuss the combination results, the combinations are categorised into two types: 1) Speech-based information combinations, 2) Music-based (vocal and instrumental) experience combinations. In the speech-based information combinations type, the combinations having both the streams from speech-based

information types (that is, monolog, interview, commentary, news headlines) were categorised, whereas, in music-based experience combinations, the combinations having a stream either from the song or music types were categorised.

Speech-based information combinations

In this category, monolog with interview, monolog with commentary, monolog with news headlines, interview with commentary, interview with new headlines, and commentary with news headlines combinations are discussed. For all of these combinations, following the analysis approach adopted in the baseline and the concurrent condition, the mean of the responses submitted by the users against rating subscales in each combination is first calculated, and then based on the means of rating subscales and the weights, the perceived workload index score for listening task in each combination is calculated.

In monolog with interview combination, the mean values of the mental demand, physical demand, temporal demand, effort, frustration, and performance were 71.75, 47.25, 55.50, 75.00, 57.00, 46.87 respectively, whereas, the index score for this combination was 67.74. Similarly, regarding the frequent and preference scales of this combination, the mean ratings was 26.62, and 23.75 respectively. Two-way ANOVA comparing the mean scores of the monolog with interview combination with the baseline condition showed that the presentation type (monolog with interview — baseline) has a significant impact on user response, $F(1, 702) = 9.71, p < 0.002$. Also, the interaction between the presentation and rating scales, $F(8, 702) = 48.125, p < 0.01$, had a significant impact on user's response. The *Post hoc* Tukey HSD test on the ANOVA results of the interaction between the presentation types and rating scales showed significant differences ($p < 0.05$) in all the scales, except, the mean index score ($p = 0.07$). Table 8.3 shows the statistical difference ($p - values$) between this combination and the baseline condition concerning each rating scale. The results showed that users' experience

in this combination was not as 'good' as in the baseline condition.

Table 8.3 : *Post hoc* Tukey HSD Analysis: (p - values) comparing mean ratings of speech-based concurrent combinations scales with the relevant baseline scales: (Signif.codes : $\leq 0.001 = ***$, $\leq 0.01 = **$, $\leq 0.05 = *$)

Combination	Ind.	Men.	Phy.	Tem.	Eff	Frus.	Per.	Fre.	Lik.
Mon. w. Int.	0.07	***	*	***	***	***	***	***	***
Mon. w. Com.	0.63	***	0.29	**	***	***	***	***	***
Mon. w. New.	0.97	**	0.58	0.17	**	***	**	***	***
Int. w Com.	0.18	***	**	**	***	***	***	***	***
Int. w New.	0.70	***	0.64	*	***	***	**	***	***
Com. w New.	0.45	***	0.12	***	***	***	***	***	***

In the monolog with commentary combination, the mean index score for the listening task was 63.19. The comparison of this combination with the baseline condition conducted with ANOVA showed that the presentation type (monolog with commentary — baseline) does not have a significant impact on user response, $F(1, 702) = 1.815, p < 0.178$. However, the interaction between the presentation and rating scales, $F(8, 702) = 32.44, p < 0.01$, had a significant impact on the user's response. As seen in the monolog with interview combination comparison, the *Post hoc* Tukey HSD test showed significant differences ($p < 0.05$) in almost all scales, except, the mean index score and the physical rating scale ($p > 0.05$). Similar to the monolog with interview combination, the results in this combination showed that users' experience was not as 'good' as noted in the baseline condition.

As mentioned in the methodology section, the new headlines in all combinations were presented intermittently with other continuous streams concurrently to the users. In monolog with news headlines, the mean index score for this listening task was 59.62. The ANOVA test on this combination showed that the presentation type (monolog with news headlines — baseline) does not have a significant impact on user response, $F(1, 702) = 3.036, p < 0.082$. However, the interaction between the presentation and rating scales, $F(8, 702) = 18.567, p < 0.01$, had a significant impact on the user's response. In the extended analysis using *Post hoc* Tukey

HSD test performing the comparison between this combination and the baseline condition, no significant difference appeared in index score ($p = 0.97$). However, probably because of the significant difference in other rating subscales ($p < 0.05$), users significantly preferred ($p < 0.05$) the baseline condition for frequent use and preference over monolog with news headlines combination.

In the interview with commentary combination, the mean index score for the listening task was 66.65. The statistical analysis ANOVA showed that the presentation type (interview with commentary — baseline) has a significant impact on user response, $F(1, 702) = 6.749, p < 0.01$. Also, the interaction between the presentation and rating scales had a significant impact on the user's response, $F(8, 702) = 42.724, p < 0.01$. As seen in the monolog with commentary combination comparison, the *Post hoc* Tukey HSD test showed significant differences ($p < 0.05$) in all the scales, except, the mean index score. Similar to monolog with commentary combination, the results in this combination showed that users' experience was not as 'good' as noted in the baseline condition.

In the interview with news headlines combination, where news headlines were presented intermittently with a continuous interview stream, the mean index score for listening this combination was 62.89. The ANOVA test on this combination showed that the presentation type (interview with news headlines — baseline) has a significant impact on user response, $F(1, 702) = 3.944, p < 0.047$. Also, the interaction between the presentation and rating scales had a significant impact on the user's response, $F(8, 702) = 25.729, p < 0.01$. In the extended analysis using *Post hoc* Tukey HSD test performing the comparison between this combination and the baseline condition, significant difference ($p < 0.05$) appeared in all the rating scales, except, the index score, and physical demand. The analysis shows that the users found this combination a challenging experience, and therefore, significantly preferred ($p < 0.05$) the baseline condition for frequent use and preference despite being provided with intermittent news headlines.

In the commentary with news headline combination, the commentary was presented with intermittent news headlines. Commentary in combination designs was used on the assumption that users usually do not pay in-depth attention to commentary types of information streams. Users mostly remain interested in gaining the gist from the match that they can get on different points, for example, the commentator becomes louder and passionate indicating that something interesting is happening on the field. Such cues may help the users to divert their attention immediately towards commentary with increased focus, else, pay attention towards the concurrent streams. In the commentary with news headline combination design, the mean index score for listening task was 64.21. The ANOVA test on this combination showed that the presentation type (commentary with news headlines — baseline) has a significant impact on the user response, $F(1, 702) = 4.711, p < 0.03$. Also, the interaction between the presentation and rating scales had a significant impact on the user's response, $F(8, 702) = 34.647, p < 0.01$. In *Post hoc* Tukey HSD test performing the comparison between this combination and the baseline condition, a significant difference ($p < 0.05$) appeared in almost all the rating scales, except, the index score, and physical demand scales ($p > 0.05$).

Music-based (Vocal and Instrumental) experience combinations

In addition to two concurrent speech-based information stream combinations, song (vocal) and instrumental (non-vocal) music streams in different combinations were included on the assumption that they would enhance user experience.

In the monolog with song combination, a song was presented with a monolog stream and the user experience was evaluated. In this combination, the mean index score for the listening task was 60.88. The two-way ANOVA test comparing this combination with the baseline condition showed no significant difference in users response concerning presentation type (monolog with song —

Table 8.4 : *Post hoc Tukey HSD Analysis: (p – values)* comparing mean ratings of music-based concurrent combinations combinations scales with the relevant baseline scales: (*Signif.codes* : $\leq 0.001 = ***, \leq 0.01 = **, \leq 0.05 = *$)

Combination	Ind.	Men.	Phy.	Tem.	Eff	Frus.	Per.	Fre.	Lik.
Mon. w. Son.	0.89	**	0.28	0.07	**	*	**	***	***
Mon. w. Mus.	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99
Int. w Son.	0.99	**	0.20	**	**	*	***	***	***
Int. w Mus.	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
Com. w Son.	1.00	0.28	0.90	*	*	*	***	**	***
Com w Mus.	1.00	0.44	0.91	0.22	0.54	0.82	0.08	0.08	*
New. w. Son.	1.00	0.99	0.80	0.97	0.82	0.84	0.38	0.12	0.12
New. w. Mus	1.00	1.00	1.00	1.00	1.00	0.98	0.93	0.39	0.19
Son w. Mus.	1.00	1.00	0.99	0.98	1.00	0.96	0.77	0.25	0.12

baseline), $F(1, 702) = 1.571, p < 0.21$. However, the interaction between the presentation and rating scales had a significant impact on the user's response, $F(8, 702) = 21.979, p < 0.01$. In the extended analysis using *Post hoc Tukey HSD* test performing the comparison between this combination and the baseline condition, no significant difference ($p > 0.05$) appeared in index score. However, frequent, and like scales were significantly different than the baseline condition ($p < 0.05$). Table 8.4 presents the statistical difference (*p – values*) between this combination and the baseline condition, and shows that though the users rating was the same for the index scale, they significantly preferred the baseline condition over monolog with song combination for frequent use and preference.

In two more combinations involving song, (that is, interview with song and commentary with song), users' responses were similar to the monolog with song combination. Statistical analysis, mentioned in Table 8.4, shows that though the users rating was the same for index scale, they significantly preferred the baseline condition over these combinations.

In the news headline with song design, intermittent news headlines were combined with song. The results showed that the mean index score for listening task was 55.60. The two-way ANOVA test comparing this combination with

the baseline condition showed no significant difference in the users' responses concerning presentation type (news headlines with song — baseline), $F(1, 702) = 0.166, p < 0.684$. In *Post hoc* Tukey HSD test performing the comparison between this combination and the baseline condition, no significant difference ($p > 0.05$) appeared in all the rating scales. The statistical analysis showed that the user experience in this combination was similar to the baseline condition.

In the monolog with music combination, results show that users have a greater interest in this combination than the combinations previously discussed. In this combination, the mean index score for the listening task was 53.38. The two-way ANOVA test comparing this combination with the baseline condition shows neither a significant difference in presentation type (monolog with song — baseline), $F(1, 702) = 0.169, p < 0.681$, nor in the interaction between the presentation and rating scales, $F(8, 702) = 1.177, p < 0.31$. Since no significant impact appeared in the interaction between the presentation and rating scales, in the extended analysis using *Post hoc* Tukey HSD test performing the comparison between this combination and the baseline condition, no significant difference ($p = 1$) appeared in all the rating scales. The statistical analysis showed that the user experience was similar to the baseline condition.

In the other four combinations involving music (instrumental — non-vocal), that is interview with music, commentary with music, news headlines with music and song with music, similar statistical results appeared as seen in monolog with music combination discussed above, see Table 8.4. There was one exception, as a significant difference appeared in like scale in commentary with music combination. This statistical analysis shows that the user experience was similar to the baseline condition in each concurrent combination involving music.

8.3.3 Information Streams Impact in Concurrent Communication

Besides discussing each combination and comparing them with the baseline condition, an overall comparison is carried out regarding each information type to determine its impacts when presented concurrently with the rest of the information types. In other words, the viability of the information types to be presented concurrently with other information streams is determined.

In the following subsections, each of the information streams is individually discussed and calculated the mean index score and mean rating for the subjective scales (as was completed in the combination analysis). For each information type, the results are compared with the baseline condition statistically, following the pattern mentioned in the combination types analysis. Before discussing each information type, Figure 8.5 is first presented to show the results for each information type, compared with the baseline condition depicted with a continuous black line. The statistical comparison of each combination with the baseline condition using *Post hoc* Tukey HSD test is mentioned in Table 8.5.

Table 8.5 : *Post hoc* Tukey HSD Analysis: (*p* – values) comparing mean ratings of each stream type with the relevant baseline scales: (*Signif.codes* : $\leq 0.001 = ***$, $\leq 0.01 = **$, $\leq 0.05 = *$)

Stream Type	Ind.	Men.	Phy.	Tem.	Eff	Frus.	Per.	Fre.	Lik.
Monolog	0.1	***	**	***	***	***	***	***	***
Interview	*	***	***	***	***	***	***	***	***
Commentary	*	***	***	***	***	***	***	***	***
News Headlines	0.46	***	*	***	***	***	***	***	***
Song	0.86	***	*	***	***	***	***	***	***
Music	1.00	0.56	0.55	0.38	0.95	0.76	0.08	*	**

This analysis starts with the monolog, and follows the same pattern adopted previously, calculating the overall mean values for index score based on the subjective subscales. The ratings for the other two scales that include frequent and like are also mentioned. The results showed that the index score for listening to a concurrent combination that had a stream type of monolog was 60.96.

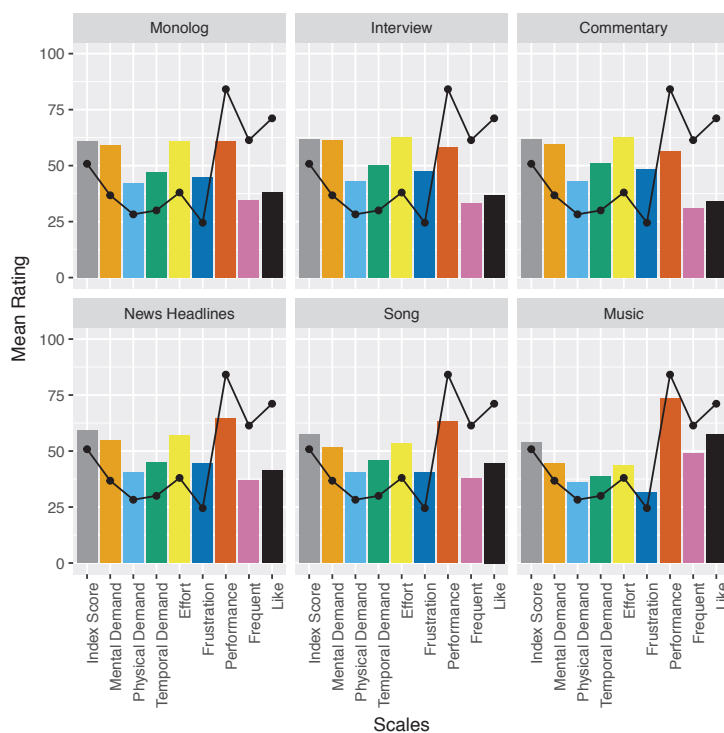


Figure 8.5 : Users' experience regarding each information type when presented with rest of the information types, and also compared with the baseline condition shown with a continuous line

In this analysis, the two way ANOVA is performed on the results and then used *Post hoc* Tukey HSD test for extended analysis to determine the significant difference between the scales of information type being discussed and the baseline condition. Each information type is compared with the baseline condition because, in the baseline condition, a randomly picked stream of any type from 6 information streams used in this study was presented to users sequentially. Therefore, an information type is compared when presented concurrently with the information types (baseline) that were presented sequentially.

For this information type, the two-way ANOVA showed that the stream type (monolog — baseline) had a significant impact on users' response $F(1, 2502) = 6.643, p < 0.01$. Also, the interaction between the presentation and rating scales, $F(8, 2502) = 55.155, p < 0.01$, had a significant impact on the users' response. Moreover, the extended analysis using *Post hoc* Tukey HSD test showed no sig-

nificant difference ($p > 0.05$) regarding mean index score. Table 8.5 shows the statistical difference ($p - values$) between this stream type and the baseline condition. The results for music, news headlines, and song, as shown in Table 8.5, appeared similar as seen in monolog type of stream. In all these types of streams, the index score difference was non-significant compared to the baseline condition, but users still preferred the baseline condition more than the concurrent types of information streams.

For the interview type, the same analysis pattern that was adopted for monolog type was followed. In this information type, the overall mean index score was 61.93. The two-way ANOVA on this information type showed that the stream type (interview — baseline) had a significant impact on the users' response $F(1, 2502) = 10.109, p < 0.001$. Also, the interaction between the presentation and rating scales, $F(8, 2502) = 62.502, p < 0.01$, had a significant impact on the users' response. The extended analysis using *Post hoc* Tukey HSD test showed significant differences ($p < 0.05$) in all the scales including mean index score ($p < 0.05$). This shows that the user experience was the least favoured in this information type when compared to the rest of the information types when presented concurrently.

In the commentary type of information stream, as shown in Table 8.5, the results appeared similar to the interview type of information stream. Therefore, the user experience was the least favoured in this information type when compared to the rest of the information types when presented concurrently.

8.3.4 Impact of Presentation in Left — Right Ears

In this study, the first type of information, that is monolog streams, was always played in the left ear for all relevant concurrent combinations. Similarly, the last information stream type, that is music, was set to play in the right ear, always. However, the rest of the information streams, in some combinations were presented in the left ear, and in some combinations in the right ear.

In this study, the interview stream was once presented in the right ear, and four times it was presented in the left ear of users. The commentary stream was presented twice in the right ear, and the remaining three times it was presented in the left ear. Regarding news headlines, it was presented in the left ear twice, and three times in the right ear. Finally, the song was presented once in the left, and the remaining four times to the right ear of the users in concurrent combinations. The composition is also indicated in Table 8.1 as mentioned in the method section above.

This composition enabled us to further extended the analysis to see the users' response with reference to presenting information in different ears, and to determine whether one ear had an advantage over the other ear in terms of enhancing the user experience, or not. For this, users response are compared regarding each information stream concerning its concurrent presentation in different ears. The analysis revealed an interesting pattern and showed that users reported lower workload index score for each of the information streams presented in the left ear, and similarly, rated higher for frequent and like scales for left ear presentation. Figure 8.6 and 8.7 show the pattern.

Following the same statistical pattern, a three-way ANOVA was performed which included the interactions between the three independent variables in determining the significant impact of all the independent variables on the user's response. The statistical results show that the ear presentation was significant $F(1, 4428) = 7.718, p = 0.005$ $F(1, 7128) = 15.989, p < 0$ in impacting the user response. However, the three-way interaction between the information type, scale type, and the ear presentation variables had a non-significant impact $F(24, 4428) = 0.392, p < 0.997$ $F(24, 7128) = 0.977, p < 0.494$ on user response, and therefore, the *Post hoc* Tukey HSD analysis on ANOVA results is not performed.

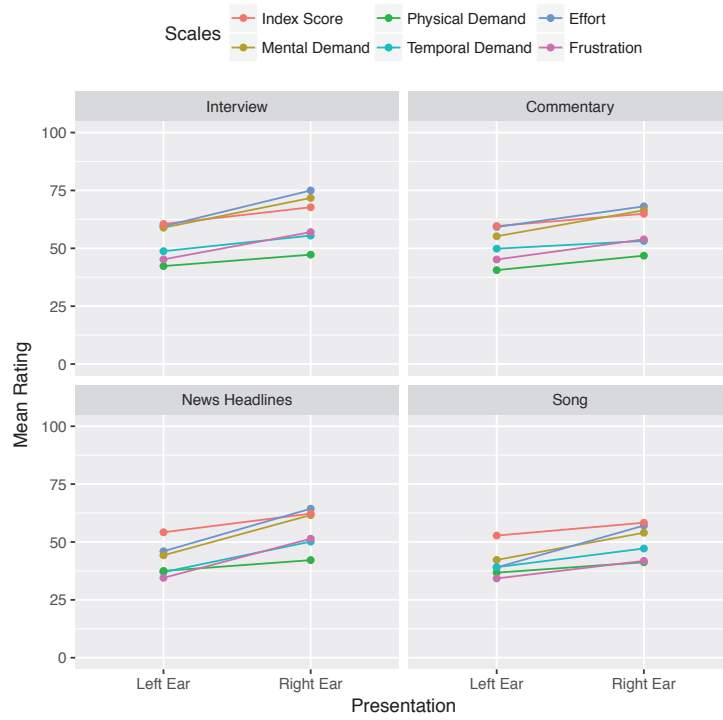


Figure 8.6 : **Impact of Users' Experience (Stress):** in each of the four information types with reference to ear presentation

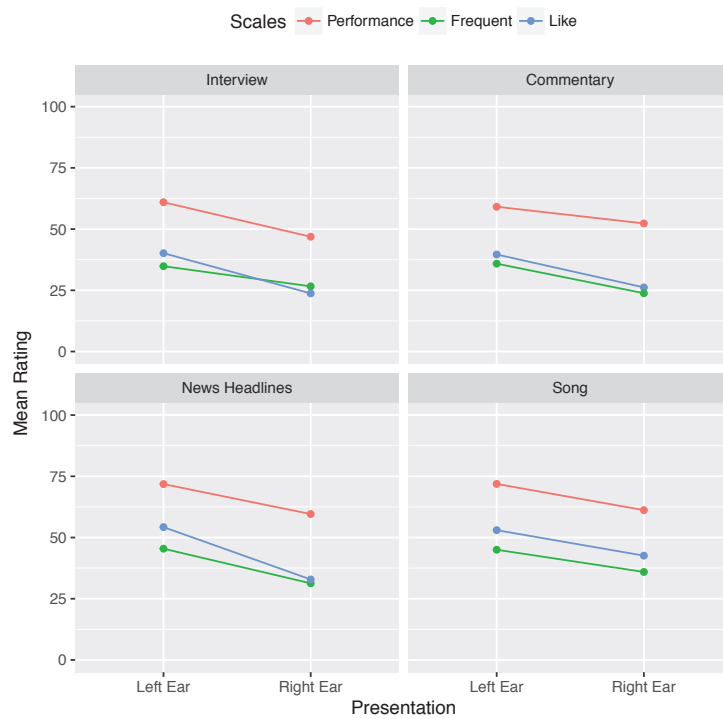


Figure 8.7 : **Impact of Users' Experience (Acceptance):** in each of the four information types with reference to ear presentation

8.4 Discussion

In this analysis, the perceived workload index score is calculated for each of the combinations and the baseline condition. The study showed that the perceived workload index was the lowest in the baseline condition. Though there is an ascending order of concurrent combinations, as shown in Figure 8.8, the statistical tests, as discussed in section 8.3, showed no significant difference ($p > 0.05$) between the baseline condition and each of the concurrent combinations. In conclusion, the perceived workload index score in concurrent and the baseline condition has no significant difference.

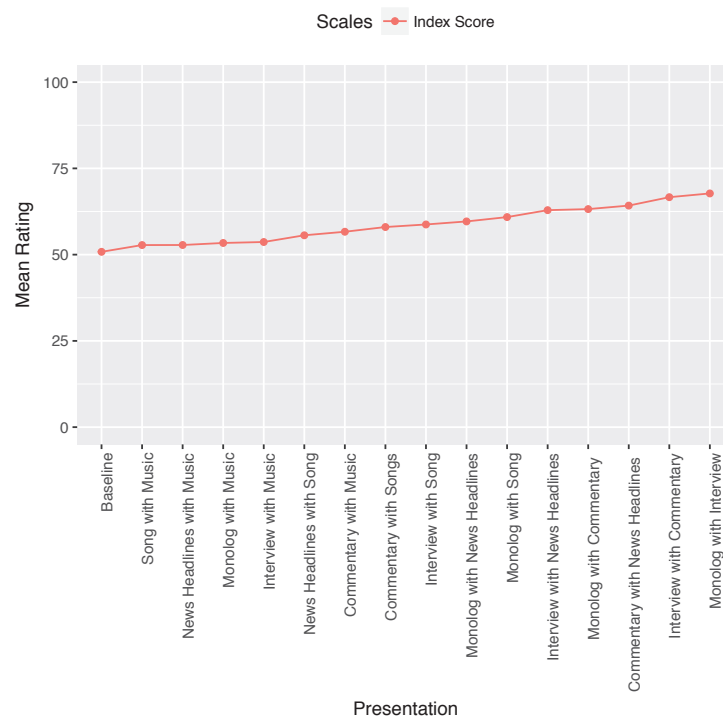


Figure 8.8 : **Perceived Workload Index Score:** an order of combinations with reference to their listening task index score reported by the users

However, contrary to the results found for perceived workload index, user responses in preference and frequently using different combinations were significantly different when concurrent conditions were compared to the baseline condition. As shown in Figure 8.9, for preference and frequent use, users again

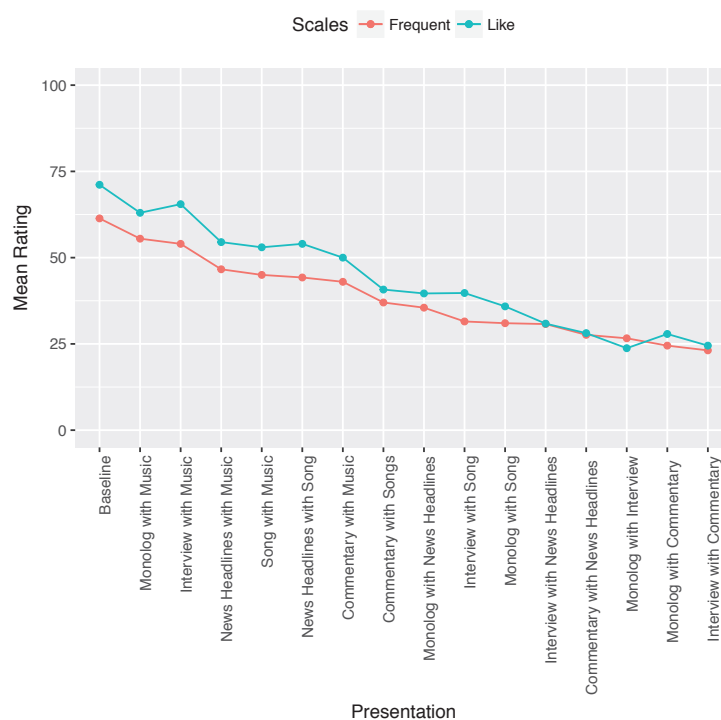


Figure 8.9 : **Ratings for Frequent and Like (Preference) Scales:** an order of combinations with reference to users' ratings regarding frequent and preference

rated the baseline condition the highest, followed by the following order of the streams with reference to frequent use. In the statistical analysis for many of the streams, users' ratings regarding frequent and like scales was significantly less than the baseline condition. The analysis shows that though the perceived workload index remains similar in baseline condition and concurrent combinations, comparing to some of the combinations, for example, interview with commentary, monolog with commentary, monolog with interview, users significantly ($p < 0.05$) preferred baseline condition in terms of preference and their likely frequent use.

The illustration in Figure 8.9, shows a relationship between frequent and preference (like) scales and looks directly proportional to each other. The analysis also shows an inverse relationship between perceived workload index score and frequent & like scales. The inversely proportional relationship between the index score and the frequent and like scales is illustrated in Figure 8.10. This relationship

shows that an increase in the perceived workload index for listening task means the relevant concurrent combination would less likely be preferred for frequent use by the users.

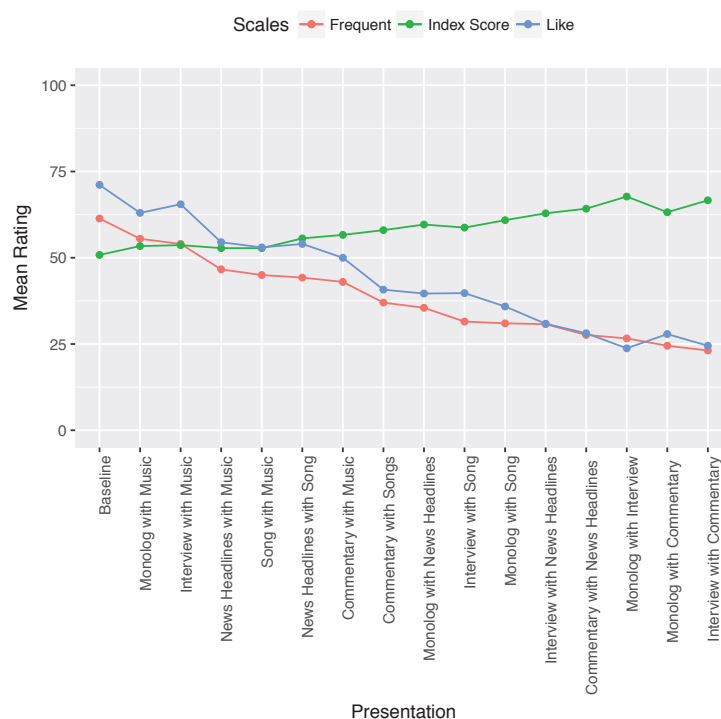


Figure 8.10 : **Order of Combinations:** with reference to their listening task index score, ratings for frequent and preference reported by the users

The pattern in Figure 8.9 also suggests that the perceived workload for a listening task in concurrent combination is dependent on the type of information as well as the amount of information presented to the users. From the order of the combinations appearing in Figure 8.9, the combinations created with music were preferred the most within concurrent combinations, followed by song-based concurrent combinations. This shows, as the music and songs usually do not require focused attention to process the information stream, there is apparently less cognitive load as users rated them high for frequent use. Similarly, in news headlines, the controlled and limited amount of information was being provided intermittently to the users in chunks, therefore, users selected it the third highest

choice to hear them in all concurrent combinations. Similarly, the concurrent combinations created with monolog, interview, and commentary were continuously delivering a high amount of voice-based information, therefore, were rated low for frequent use by the users. This pattern shows that the high amount of information delivery requiring greater attention and cognitive processing from the users to comprehend information makes it less acceptable for the users.

The extended analysis discussing the viability of information type in concurrent combination also validates the above observation that the perceived workload for a listening task in concurrent combination is dependent on the type of information as well as the amount of information presented to the users. The order is shown in Figure 8.11 validates the same as users rated music the highest regarding frequent use when listened in concurrent combination, followed by song as the second.

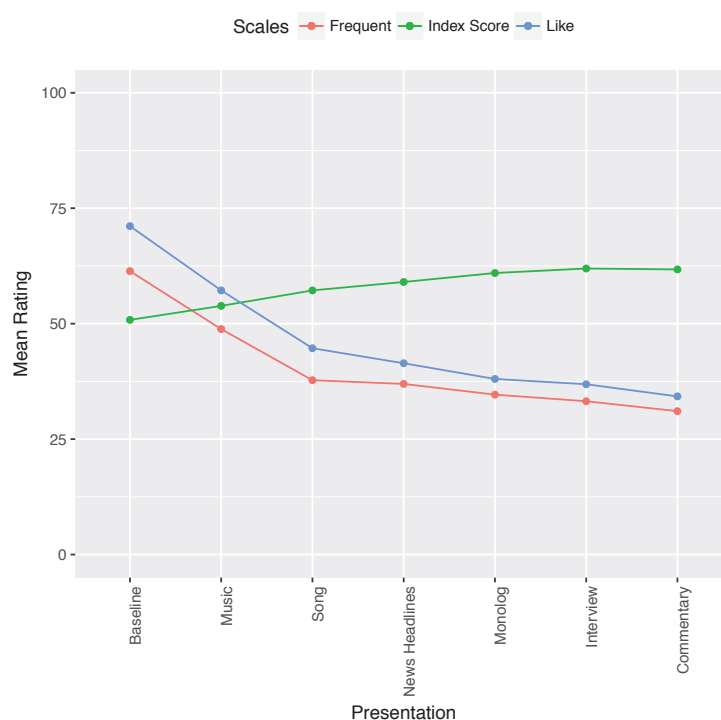


Figure 8.11 : **Order of Information Types** with reference to their potential for being a part of concurrent communication

From the information type providing speech-based information (non-music/song),

users rated news headlines the highest for frequent use when listening to concurrent combinations. Rating news headlines the highest support previous studies of this research which show the intermittent design with the spatial difference in sources is the best form to communicate multiple information concurrently. In the previous study, in the intermittent form of concurrent communication, users comprehended the content equal to the amount that they comprehended in the baseline sequential presentation. In this study, users reported that intermittent form of communication creates the least perceived workload index in speech-based information communication.

Moreover, from the speech-based information types, after the news headlines, users rated monolog the second highest information stream regarding frequent use. In monolog, one speaker presented information, whereas in interview (dialog) two speakers were involved in presenting information. Comparing these two types of information streams, users rated monolog higher than the interview. This shows that the number of talkers in concurrent streams also affects users perceived workload index. A stream with one talker is rated higher in terms of frequent use than the stream involving two talkers.

Moreover, users rated commentary the least. There could be many factors, such as there being background noise generated from the spectators in the commentary stream or the speaking speed of the commentator being fast in order to keep up with the pace of the game. The higher speaking rate/pace might also have created a difference in monolog and dialog as the speaking rate in monolog was less than the dialog.

As discussed in the ear impact section, the detailed analysis showed that in concurrent communication, users preferred an information type more when presented in the left ear as compared to the right ear presentation. In all the information types, the analysis showed that users reported lower workload index score for each of the information stream when it was presented in the left ear, and

rated higher for frequent and preference scales compared to right ear presentation, and vice-versa.

8.5 Limitations and Future Work

Though this study tried to compare different combination streams comprehensively, the impact on user experience by the content played and user's interest in the topics of the information streams cannot be fully overruled.

In the analysis of the combination types, it appeared that the monolog presented with the music achieved almost the same user experience as users enjoyed in the baseline condition. The user experience almost remained the same in baseline condition and combinations with music. This shows that presenting music with an information stream does not create a significant difference in user experience as compared to the baseline condition. Some users rated combinations with music higher than the baseline condition in terms of frequent use and preference. It sets another direction for investigation to test the impact of music presented in one ear while comprehending content from other voice-based streams presented in the other ear. Many people do tasks with music playing in the background. Dichotic listening could be tested where a speech-based information stream is provided in one ear and the music stream in another ear. A study based on the same design pattern that was adopted in Study 3 (see Chapters 6, 7) could be used to compare the content comprehension in such concurrent communication with the baseline sequential communication.

The results of this recent study discussed in this chapter are submitted in following high impact Journal:

Publication 8: —, “Investigating cognitive workload in concurrent speech-based information communication,” *The Journal of the Acoustical Society of America (JASA)*, vol. -, no. -, pp. 1–20, 2019, submitted.
Attached as Appendix-N.

Chapter 9

Conclusion

In this extensive research of four user studies regarding concurrent information communication, various findings were analysed. In the following points, the study-wide findings of this research are briefly discussed:

- Chapter 3 reviewed a small investigation, that explored whether voice-based multiple information communication in concurrent form is possible in Human-Computer Interaction. Users showed an interest in concurrent information communication and were able to both discriminate and understand the voice streams using selection and attention abilities. Users were able to receive multiple information streams meaningfully in less time. The results of Study 1 encouraged the exploration of concurrent information communication design further.
- Chapter 4 reported on an investigation undertaken with both Visually Challenged Users (VCU) and Sighted Users (SU), users found the continuous form of delivery more appropriate than the interval-based method. However, some of the users expressed that dichotic audio technique is helpful in segregating multiple voice streams from each other. This study suggested that spatial difference between the streams presented needs to be explored further along with player controls. Based on this investigation, a framework is introduced for concurrent information communication and a web-based prototype *Vinformize* is developed to communicate multiple information streams to users concurrently. It is expected that the application of this new framework to information systems that provide multiple concurrent

communication will provide a better user experience for users subject to their contextual and perceptual needs and limitations.

- Chapters 5, 6 and 7, reported that communicating multiple speech-based information streams is equally plausible using the high playback-rate and concurrent approaches. However, the performance in comprehension will be significantly lower in these approaches compared to the baseline condition. The empirical study showed that the concurrent speech-based information designs involving intermittent form and a spatial difference in sources of the streams provide satisfying comprehensibility. In addition, analysis of the study showed: a) In the concurrent speech-based information communication, users successfully answered the main questions, some of the implied questions, as well as the questions that required detailed information. b) Concurrent speech-based information communication works better when the information in stereo audio quality is provided intermittently and dichotically. c) The Diotic-Monotic design involving REA does not serve an advantage in speech-based concurrent communication. d) The high-frequency voice attracts more attention from listeners where there are competing voices. e) Male users were more interested in speech-based concurrent communication compared to female users. f) The comprehension pattern remains similar in concurrent speech-based communication, as seen in sequential communication. g) Besides encouraging results in concurrent speech-based information communication, users also reported a high cognitive load.
- In chapter 8 reported that: a) Cognitive workload in concurrent and the baseline condition has no significant difference b) Opposing the cognitive workload index, users response in preferring and frequently using different combinations remains significantly low compared to the baseline condition c) Listening tasks in concurrent combination are dependent on the type of

information as well as the amount of information presented to users d) Combinations created with music were liked the most in concurrent combinations, followed by song e) From the information types providing speech-based information (non-music/song), users rated intermittent news headlines the highest regarding frequent use f) The intermittent form of communication creates the lowest cognitive workload in speech-based information communication g) Users preferred monolog over interview (dialog) showing that streams with one talker receive higher rating in terms of frequent use than the stream involving two talkers h) Users rated commentary the lowest as there was background noise within the stream, and also the speaking speed of the commentator was high showing that higher speaking rate/pace might also have had an impact i) In concurrent communication, users like an information type more when presented in the left ear as compared to the right ear presentation j) Monolog presented with music dichotically achieved almost the same user experience as users had in the baseline condition, inviting another investigation to test the impact of music while comprehending the content from the other speech-based stream presented in another ear.

Overall, this research work contributes to pursuing the 'design for all paradigm' that aims to enable sighted and the unsighted persons to concurrently interact with digital information applications, e.g., 1) Listening to information streams, 2) Finding relevant information items, 3) Scanning for specific information, 4) Notifications using a secondary audio channel, 5) Multitouch to multi-sound, 6) Collaborative environments, 7) Feedback for text-entry corrections, and 8) Give support to shortcut navigation via lists of links and headings etc. (Guerreiro, 2016b).

Our research carefully explored different angles of concurrent communication by taking users' interest and their expectations from such systems into account. Users showed an interest in concurrent information communication and were

able to both discriminate and understand the voice streams using selection and attention abilities. However, they reported about the high cognitive load while listening to concurrent speech-based information. Users asked to provide them with authority to decide about listening to information sequentially or concurrently depending on their information seeking context. Users reported that the spatial difference was one of the most critical factors for segregating competing streams. Therefore, panning is proposed to be an integral control that must be provided to users for segregating competing streams and comprehending concurrent information efficiently. Also, users reported that high playback-rate has the potential to be used for communicating multiple information streams efficiently. Therefore, it is also proposed to provide users with playback-rate control along with the panning and other standard audio controls. The proposed framework would help users to seek information according to their information seeking context.

Our research shows that concurrent comprehension improves when additional design factors involved in concurrent speech-based designs. The secondary information stream provided intermittently in speech-based concurrent design having stereo-channelled audio quality renders better concurrent-speech comprehension compared to the continuous presentation. In addition to these factors, concurrent presentation further gets facilitated with a spatial separation cue that helps users to attain similar comprehensibility that they achieve in the baseline condition. In such a design, the information comprehension remains similar not only in main but also in implied and detailed content. The best performance in the intermittent concurrent design provides opportunities to the voice-based human-computer interaction designers for introducing the same strategy that can be seen in GUI using overlay or lightbox, discussed in section 1.4 in Chapter 1.

Our research also measured the cognitive workload while listening to the concurrent information streams and reports that the perceived workload index score in concurrent and the baseline condition has no significant difference. However,

users response in preferring and frequently using different combinations remains significantly different compared to the baseline condition. Combinations created with music were preferred the most in concurrent combinations, followed by song. Some users rated combinations with music higher than the baseline condition regarding frequent use and preference. This invites another investigation to test the impact on comprehension by user's custom-selected music presented in one ear and listening to the user's custom selected content in the other ear. A study based on the same design pattern adopted in Study 3 (Chapters 5, 6, and 7) can be used to compare the content comprehension in such concurrent communication. Additionally, regarding the future work, the designed framework and the prototype developed during this research may help in studying user's behaviour *In-the-Wild Usage* (Guerreiro, 2016b) of concurrent speech-based information communication. One more important aspect for investigation could be to see whether users' frequent interaction with concurrent information enhances their abilities to comprehend information, and as well as the interest in "Concurrent Information Communication in Voice-based Interaction."

Bibliography

- S. K. Agarwal, A. Jain, A. Kumar, A. A. Nanavati, and N. Rajput, "The Spoken Web: A Web for the Underprivileged," *ACM SIGWEB Newsletter*, no. Summer, pp. 1–9, jun 2010.
- S. Akram, "Neural and Computational Approaches to Auditory Scene Analysis," Ph.D. dissertation, University of Maryland, 2015.
- T. L. Arbogast and G. Kidd, "Evidence for spatial tuning in informational masking using the probe-signal method," *The Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1803–1810, 2000.
- S. Arlinger, T. Lunner, B. Lyxell, and M. K. Pichora-fuller, "The emergence of cognitive hearing science," *Scandinavian Journal of Psychology*, vol. 5, no. 50, pp. 371–384, 2009.
- B. Arons, "Efficient listening with two ears: Dichotic time compression and spatialization," in *Proceedings of the 2nd International Conference on Auditory Display*. Georgia Institute of Technology, 1994, pp. 171–178.
- S. Asthana, P. Singh, and A. Singh, "Mocktell: Exploring challenges of user emulation in interactive voice response testing," in *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering*. ACM, 2013, pp. 427–428.
- S. Asthana, P. Singh, and A. Singh, "A usability study of adaptive interfaces for

- interactive voice response system," in *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 2013, pp. 1–2.
- Audacity, "Audacity," <https://www.audacityteam.org/>, [Online; accessed 23-Dec-2018].
- J. Aydelott, D. Baer-Henney, M. Trzaskowski, R. Leech, and F. Dick, "Sentence comprehension in competing speech: Dichotic sentence-word priming reveals hemispheric differences in auditory semantic processing," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 1108–1144, 2012.
- J. Aydelott, Z. Jamaluddin, and S. Nixon Pearce, "Semantic processing of unattended speech in dichotic listening," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 964–975, 2015.
- A. Baird, S. H. Jørgensen, E. Parada-Cabaleiro, S. Hantke, N. Cummins, and B. Schuller, "Perception of Paralinguistic Traits in Synthesized Voices," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*. ACM, 2017, pp. 1–5.
- M. Bakhouya and J. Gaber, "Service Composition Approaches for Ubiquitous and Pervasive Computing Environments: A Survey," in *Agent Systems in Electronic Business*. IGI Global, 2007, pp. 323–350.
- G. Baruah, M. D. Smucker, and C. L. Clarke, "Evaluating Streams of Evolving News Events," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015, pp. 675–684.
- BBC, "BBC-URDU," <http://www.bbc.com/urdu>, 2016, [Online; accessed 01-May-2016].
- D. Beattie, L. Baillie, and M. Halvey, "A comparison of artificial driving sounds for

- automated vehicles,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 451–462.
- D. Beattie, L. Baillie, and M. Halvey, “Exploring How Drivers Perceive Spatial Earcons in Automated Vehicles,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–24, 2017.
- V. Best, F. J. Gallun, A. Ihlefeld, and B. G. Shinn-Cunningham, “The influence of spatial separation on divided listening,” *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1506–1516, 2006.
- K. Biatov and J. Koehler, “An audio stream classification and optimal segmentation for multimedia applications,” in *Proceedings of the 11th ACM International Conference on Multimedia*. ACM, 2003, pp. 211–214.
- J. K. Bizley and Y. E. Cohen, “The what, where and how of auditory-object perception,” *Nature Reviews Neuroscience*, vol. 14, no. 10, pp. 693–707, 2013.
- L. Brayda, F. Traverso, L. Giuliani, F. Diotalevi, S. Repetto, S. Sansalone, A. Trucco, and G. Sandini, “Spatially selective binaural hearing aids,” in *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 957–962.
- E. Brazil and M. Fernström, “Investigating concurrent auditory icon recognition,” in *Proceedings of the 12th International Conference on Auditory Display (ICAD)*. Georgia Institute of Technology, 2006, pp. 51–58.
- E. Brazil, M. Fernstrom, and J. Bowers, “Exploring concurrent auditory icon recognition,” in *Proceedings of the 15th International Conference on Auditory Display (ICAD)*. Georgia Institute of Technology, 2009, pp. 1–4.

- A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.
- D. E. Broadbent, *The effects of noise on behaviour*. Pergamon Press, 1958.
- D. Brock, B. McClimens, J. G. Trafton, M. McCurry, and D. Perzanowski, "Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech," in *Proceedings of the 14th International Conference on Auditory Display (ICAD)*. Georgia Institute of Technology, 2008, pp. 1–8.
- D. Brock, C. Wasylyshyn, B. McClimens, and D. Perzanowski, "Facilitating the watchstander's voice communications task in future navy operations," in *MIL-COM 2011 Military Communications Conference*, 11 2011, pp. 2222–2226.
- D. Brock, C. Wasylyshyn, and B. McClimens, "Word spotting in a multichannel virtual auditory display at normal and accelerated rates of speech," in *Proceedings of the 22nd International Conference on Auditory Display (ICAD)*. Georgia Institute of Technology, 2016, pp. 130–135.
- Brokx and Nootboom, "Intonation and the perceptual separation of simultaneous voices," *Journal of Phonetics*, vol. 10, no. 1, pp. 23–36, 1982.
- A. W. Bronkhorst and R. Plomp, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3132–3139, 1992.
- D. S. Brungart and B. D. Simpson, "Optimizing the spatial configuration of a seven-talker speech display," *ACM Transactions on Applied Perception*, vol. 2, no. 4, pp. 430–436, 2005.
- D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott, "Informational and

- energetic masking effects in the perception of multiple simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2527–2538, 2001.
- M. Bryden, "An overview of the dichotic listening procedure and its relation to cerebral organization." 1988.
- D. Cabrera, S. Ferguson, and G. Laing, "Development of auditory alerts for air traffic control consoles," in *119th Convention of the Audio Engineering Society*. Audio Engineering Society, 2005, pp. 2–21.
- R. L. Canosa, "Real-world vision: Selective perception and task," *ACM Transactions on Applied Perception*, vol. 6, no. 2, pp. 1–34, 2009.
- R. Carhart, T. W. Tillman, and E. S. Greetis, "Perceptual Masking in Multiple Sound Backgrounds," *The Journal of the Acoustical Society of America*, vol. 45, no. 3, pp. 694–703, 1969.
- S. Carlile and D. Schonstein, "Frequency bandwidth and multi-talker environments," in *120th Convention of Audio Engineering Society*. Audio Engineering Society, 2006, pp. 1–8.
- R. P. Carlyon, C. J. Plack, D. A. Fantini, and R. Cusack, "Cross-modal and non-sensory influences on auditory streaming," *Perception*, vol. 32, no. 11, pp. 1393–1402, 2003.
- G. Chernyshov, B. Tag, J. Chen, V. Noriyasu, P. Lukowicz, and K. Kunze, "Wearable ambient sound display," in *Proceedings of the 2016 ACM International Symposium on Wearable Computers*. ACM, 2016, pp. 58–59.
- E. C. Cherry and W. K. Taylor, "Some Further Experiments upon the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 26, no. 4, pp. 554–559, 1954.

- d. A. Cheveigné, S. McAdams, J. Laroche, and M. Rosenberg, "Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3736–3748, 1995.
- K. Church, M. Cherubini, and N. Oliver, "A large-scale study of daily information needs captured in situ," *ACM Transactions on Computer-Human Interaction*, vol. 21, no. 2, pp. 1–46, 2014.
- A. Claude, "Breaking the wave: Effects of attention and learning on concurrent sound perception," *Hearing Research*, vol. 229, no. 1–2, pp. 225–236, 2007.
- B. M. Clopton and F. A. Spelman, *Auditory system*. CRC Press & IEEE Press Boca Raton, 1995.
- A. R. A. Conway, N. Cowan, and M. F. Bunting, "The Cocktail Party Phenomenon Revisited," *Psychonomic Bulletin & Review*, vol. 8, no. 2, pp. 331–335, 2001.
- M. D. Copenhaver and B. S. Holland, "Multiple comparisons of simple effects in the two-way analysis of variance with fixed effects," *Journal of Statistical Computation and Simulation*, pp. 1–15, 1988.
- N. Cowan, *Attention and memory: An integrated framework*. Oxford University Press, 1998.
- R. Cusack, J. Deeks, G. Aikman, and R. P. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 4, pp. 643–656, 2004.
- R. J. P. Damaceno, J. C. Braga, and J. P. Mena-Chalco, "Mobile device accessibility for the visually impaired: problems mapping and recommendations," *Universal Access in the Information Society*, vol. 17, no. 2, pp. 421–435, 2018.

- C. Darwin and R. Carlyon, *Hearing*. Academic Press, 1995.
- M. L. D. Deroche and J. F. Culling, "Voice segregation by difference in fundamental frequency: Effect of masker type," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. EL465–EL470, 2013.
- M. L. D. Deroche, J. F. Culling, M. Chatterjee, and C. J. Limb, "Roles of the target and masker fundamental frequencies in voice segregation," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1225–1236, 2014.
- S. Deshpande, "On-Display Spatial Audio for Multiple Applications on Large Displays," in *Proceedings of the 2nd ACM International Workshop on Immersive Media Experiences*. ACM, 2014, pp. 19–22.
- J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations." *Psychological Review*, vol. 70, no. 1, p. 80, 1963.
- A. Dix, *Human-computer interaction*. Pearson, 2003.
- J. Doherty, K. Curran, and P. Mckeivitt, "A self-similarity approach to repairing large dropouts of streamed music," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 9, no. 3, pp. 1–20, 2013.
- K. Drossos, A. Floros, and N. Grigorios, "Affective Acoustic Ecology : Towards Emotionally Enhanced Sound Events," in *Proceedings of the 7th Audio Mostly Conference*. ACM, 2012, pp. 109–116.
- J. L. Drury, J. Scholtz, and D. Kieras, "Adapting GOMS to model human-robot interaction," in *Proceeding of the ACM/IEEE International Conference on Human Robot Interaction*. ACM, 2007, pp. 41–48.
- H. Duifhuis, L. F. Willems, and R. J. Sluyter, "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception," *The Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1568–80, 1982.

- M. Elhilali, "Modeling the Cocktail Party Problem," *Current Biology*, vol. 60, no. 22, pp. 111–135, 2017.
- M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3751–3771, 2008.
- M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLOS Biology*, vol. 7, no. 6, p. e1000129, 2009.
- M. W. Eysenck, *Psychology: An international perspective*. Taylor & Francis, 2004.
- M. W. Eysenck and M. T. Keane, *Cognitive psychology: A student's handbook*. Psychology press, 2013.
- M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, "Vinfomize: A framework for multiple voice-based information communication," in *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*. ACM, 2019, pp. 143–147.
- M. A. u. Fazal and M. Shuaib Karim, "Multiple information communication in voice-based interaction," in *Advances in Intelligent Systems and Computing*. Springer, pp. 101–111.
- M. A. u. Fazal, S. Ferguson, and A. Johnston, "Investigating Concurrent Speech-based Designs for Information Communication," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ACM. New York, NY, USA: ACM, 2018, pp. 1–8.
- M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, "Concurrent Voice-Based Multiple Information Communication: A Study Report of Profile-Based

- Users' Interaction," in *145th Convention of the Audio Engineering Society*. Audio Engineering Society, 2018.
- M. A. u. Fazal and M. Shuaib Karim, "Multiple information communication in voice-based interaction," in *Advances in Intelligent Systems and Computing*. Springer, pp. 101–111.
- M. A. u. Fazal, S. Ferguson, and A. Johnston, "Evaluation of Information Comprehension in Speech-based Designs for Concurrent Audio Streams," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. -, no. -, pp. 1–18, 2018, submitted.
- , "Investigating cognitive workload in concurrent speech-based information communication," *The Journal of the Acoustical Society of America (JASA)*, vol. -, no. -, pp. 1–20, 2019, submitted.
- , "Investigating Concurrent Speech-based Designs for Efficient Information Communication - Extended Analysis," *Journal of the Audio Engineering Society (JAES)*, vol. -, no. -, pp. 1–8, 2019, submitted.
- , "Investigating Efficient Speech-based Information Communication - A Comparison between the High-rate and the Concurrent Playback Designs," *Journal on Multimodal User Interfaces (JMUI)*, vol. -, no. -, pp. 1–8, 2019, submitted.
- W.-c. Feng, "Streaming media evolution: where to now?" in *Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video*. ACM, 2012, pp. 57–58.
- S. Ferguson, "Exploratory sound analysis: statistical sonifications for the investigation of sound," Ph.D. dissertation, University of Sydney, 2009.
- S. Ferguson and D. Cabrera, "Exploratory sound analysis: sonifying data about sound." International Community for Auditory Display, 2008, pp. 1–8.

- J. M. Festen, "Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice," *Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1295–1300, 1993.
- C. Frauenberger and T. Stockman, "Patterns in auditory menu design," in *Proceedings of the 12th International Conference on Auditory Display*. Georgia Institute of Technology, 2006, pp. 141–147.
- R. L. Freyman, U. Balakrishnan, and K. S. Helfer, "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2246–2256, 2004.
- Geo-News, "Geo-News," <http://www.geo.tv>, [Online; accessed 01-May-2016].
- S. Getzmann, E. J. Golob, and E. Wascher, "Focused and divided attention in a simulated cocktail-party situation: ERP evidence from younger and older adults," *Neurobiology of Aging*, vol. 41, pp. 138–149, 2016.
- J. Guerreiro and D. Goncalves, "Text-to-speeches: Evaluating the perception of concurrent speech by blind people," in *ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 2014, pp. 169–176.
- , "Scanning for digital content: How blind and sighted people perceive concurrent speech," *ACM Transactions on Accessible Computing*, vol. 8, no. 1, 2016.
- J. Guerreiro, "Enhancing blind people's information scanning with concurrent speech," Ph.D. dissertation, University of Lisbon, 2016.
- , "Using simultaneous audio sources to speed-up blind people's web scanning," in *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM, 2013, pp. 1–2.

- , “Towards screen readers with concurrent speech: where to go next?” *SIGACCESS Accessibility and Computing*, no. 114, pp. 12–19, 2016.
- J. Guerreiro and D. Goncalves, “Faster text-to-speeches: Enhancing blind people’s information scanning with faster concurrent speech,” in *Proceedings of the 17th International ACM - SIGACCESS Conference on Computers & Accessibility*. ACM, 2015, pp. 3–11.
- S. R. Gulliver and G. Ghinea, “Defining user perception of distributed multimedia quality,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 2, no. 4, pp. 241–257, 2006.
- A. Gutschalk, C. Micheyl, and A. J. Oxenham, “Neural correlates of auditory perceptual awareness under informational masking,” *PLOS Biology*, vol. 6, no. 6, pp. 1156–1165, 2008.
- S. G. Hart and L. E. Stavenland, “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research,” in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds. Elsevier, 1988, ch. 7, pp. 139–183. [Online]. Available: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000004342_1999205624.pdf
- R. Hausmann, L. D. Tyson, Y. Bekhouche, and S. Zahidi, “The global gender gap report.” World Economic Forum, 2012, pp. 1–381.
- M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, “Speech intelligibility and localization in a multi-source environment,” *The Journal of the Acoustical Society of America*, vol. 105, no. 6, pp. 3436–3448, 1999.
- M. L. Hawley, R. Y. Litovsky, and J. F. Culling, “The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer,” *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004.

- C. Hayward, "Listening to the Earth sing," in *Auditory Display: Sonification, Audification, and Auditory Interfaces*, vol. 18. Addison-Wesley Publishing co., 1994, pp. 369–404.
- M. A. Hearst, "'Natural' search user interfaces," *Communications of the ACM*, vol. 54, no. 11, p. 60, 2011.
- K. S. Helfer and R. L. Freyman, "Lexical and indexical cues in masking by competing speech," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 447–456, 2009.
- T. Hermann, "Taxonomy and definitions for sonification and auditory display," in *Proceedings of the 14th International Conference on Auditory Display (ICAD)*. Georgia Institute of Technology, 2008, pp. 1–8.
- A. F. Hinde, "Concurrency in auditory displays for connected television," Ph.D. dissertation, University of York, 2016.
- A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "Perceived Audio Quality for Streaming Stereo Music," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 1173–1176.
- K. Hugdahl, "Dichotic listening and attention: the legacy of Phil Bryden," *Laterality*, vol. 21, no. 4-6, pp. 433–454, 2016.
- L. E. Humes and J. R. Dubno, "Factors affecting speech understanding in older adults," in *The Aging Auditory System*. Springer, 2010, pp. 211–257.
- A. Hussain, M. A. u. Fazal, and M. S. Karim, "Intra-domain user model for content adaptation," in *Smart Innovation, Systems and Technologies*. Springer, 2015, pp. 285–295.
- I. Hussain, L. Chen, H. T. Mirza, A. Majid, and G. Chen, "Hybrid auditory feedback: A new method for mobility assistance of the visually impaired," in *Pro-*

ceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility, 2012, pp. 255–256.

- I. Hussain, L. Chen, H. T. Mirza, G. Chen, and S.-U. Hassan, “Right mix of speech and non-speech: hybrid auditory feedback in mobility assistance of the visually impaired,” *Universal Access in the Information Society*, vol. 14, no. 4, pp. 527–536, Nov 2015.
- A. Ihlefeld and B. Shinn-Cunningham, “Spatial release from energetic and informational masking in a selective speech identification task,” *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4369–4379, 2008.
- Y. Ikei, H. Yamazaki, K. Hirota, and M. Hirose, “vCocktail: multiplexed-voice menu presentation method for wearable computers,” in *Virtual Reality Conference*. IEEE, 2006, pp. 183–190.
- N. Iyer, D. S. Brungart, and B. D. Simpson, “Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task,” *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 2998–3010, 2010.
- N. Iyer, E. R. Thompson, B. D. Simpson, D. Brungart, and V. Summers, “Exploring auditory gist: Comprehension of two dichotic, simultaneously presented stories,” in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. Acoustical Society of America, 2013, pp. 050 158–050 158.
- P. J. James, S. Krishnan, and J. Aydelott, “Working memory predicts semantic comprehension in dichotic listening in older adults,” *Cognition*, vol. 133, no. 1, pp. 32–42, 2014.
- F. Jay and S. Gordon, *Cognitive Science: An Introduction to the Study of Mind*. Sage, 2005.

- M. J. Kane, A. R. A. Conway, D. Z. Hambrick, and R. W. Engle, "Variation in working memory capacity as variation in executive control," *Variation from Inter- and Intra-individual Differences*, vol. 1, pp. 21–48, 2001.
- C.-t. Kao, Y.-t. Liu, and A. Hsu, "Speeda : Adaptive Speed-up for Lecture Videos," in *Proceedings of the Adjunct Publication of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2014, pp. 97–98.
- A. Kaur and D. Dani, "Comparing and evaluating the effectiveness of mobile Web adequacy evaluation tools," *Universal Access in the Information Society*, vol. 16, no. 2, pp. 411–424, 2017.
- T. Kawashima and T. Sato, "Perceptual limits in a simulated "Cocktail party"," *Attention, Perception, and Psychophysics*, vol. 77, no. 6, pp. 2108–2120, 2015.
- R. Kazhamiakin, P. Bertoli, M. Paolucci, M. Pistore, and M. Wagner, "Having services "yourway!": towards user-centric composition of mobile services," in *Future Internet Symposium*. Springer, 2008, pp. 94–106.
- G. Kidd, C. R. Mason, T. L. Rohtla, and P. S. Deliwala, "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 422–431, 1998.
- S. Kim, "Investigating Everyday Information Behavior of Using Ambient Displays: A Case of Indoor Air Quality Monitors," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. ACM, 2018, pp. 249–252.
- J. Koehnke and J. M. Besing, "A procedure for testing speech intelligibility in a virtual listening environment," *Ear and Hearing*, vol. 17, no. 3, pp. 211–217, 1996.
- W. Köhler, *Gestalt psychology: An introduction to new concepts in modern psychology*. Liveright Publishing Corporation., 1947.

- N. Kopčo and B. G. Shinn-Cunningham, "Spatial unmasking of nearby pure-tone targets in a simulated anechoic environment," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2856–2870, 2003.
- P. Kortum, *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other*. Morgan Kaufmann Publishers Inc., 2008.
- G. Kramer, "An introduction to auditory display in kramer, g.(ed.) auditory display," 1994.
- K. Kummamuru, A. Jujuru, and M. Duggirala, "Generating facets for phone-based navigation of structured data," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, 2012, pp. 1283–1292.
- P. Larsson, "Speech Feedback Reduces Driver Distraction Caused by In-vehicle Visual Interfaces," in *Proceedings of the Audio Mostly 2016*. ACM, 2016, pp. 7–11.
- E. A. Lawson, "Decisions concerning the rejected channel." *The Quarterly Journal of Experimental Psychology*, vol. 18, no. 3, pp. 260–265, 1966.
- G.-p. Li and G.-y. Huang, "The 'Core-Periphery' pattern of the globalization of electronic commerce," in *Proceedings of the 7th International Conference on Electronic Commerce*. ACM, 2005, pp. 66–69.
- D. Liang, Y. Xiao, Y. Feng, and Y. Yan, "The role of auditory feedback in speech production: Implications for speech perception in the hearing impaired," in *Proceedings of the 14th International Symposium on Integrated Circuits*, 2015, pp. 192–195.
- P. Lindborg and N. Kwan, "Audio quality moderates localisation accuracy: Two distinct perceptual effects?" in *138th Convention of the Audio Engineering Society*, vol. 2. Audio Engineering Society, 2015, pp. 797–806.

- R. Y. Litovsky, B. J. Fligor, and M. J. Tramo, "Functional role of the human inferior colliculus in binaural hearing," *Hearing Research*, vol. 165, no. 1-2, pp. 177–188, 2002.
- R. H. Lorenz, A. Berndt, and R. Groh, "Designing auditory pointers," in *Proceedings of the 8th Audio Mostly Conference on - AM '13*. ACM, 2013, pp. 1–6.
- R. A. Lutfi, "How much masking is informational masking?" *The Journal of the Acoustical Society of America*, vol. 88, no. 6, pp. 2607–2610, 1990.
- N. W. MacKeith and R. R. Coles, "Binaural advantages in hearing of speech," *The Journal of Laryngology & Otology*, vol. 85, no. 3, pp. 213–232, 1971.
- A. Matassa, L. Console, L. Angelini, M. Caon, and O. A. Khaled, "Workshop on full-body and multisensory experience in ubiquitous interaction," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers - UbiComp '15*. ACM, 2015, pp. 923–926.
- D. K. McGookin and S. A. Brewster, "An investigation into the identification of concurrently presented earcons," in *Proceedings of 2003 International Conference on Auditory Display (ICAD)*. Georgia Institute of Technology, 2003, pp. 42–46.
- , "Understanding concurrent earcons: applying auditory scene analysis principles to concurrent earcon recognition," *ACM Transactions on Applied Perception*, vol. 1, no. 2, pp. 130–155, 2004.
- I. Medhi, S. Patnaik, E. Brunskill, S. N. Gautama, W. Thies, and K. Toyama, "Designing mobile interfaces for novice and low-literacy users," *ACM Transactions on Computer-Human Interaction*, vol. 18, no. 1, pp. 1–28, 2011.
- C. Micheyl, R. P. Carlyon, A. Gutschalk, J. R. Melcher, A. J. Oxenham, J. P. Rauschecker, B. Tian, and E. C. Wilson, "Role of Auditory Cortex in the For-

- mation of Auditory Objects," *Hearing Research*, vol. 229, no. 1-2, pp. 116–131, 2007.
- G. A. Miller, "The masking of speech." *Psychological Bulletin*, vol. 44, no. 2, p. 105, 1947.
- R. G. Miller, *Simultaneous Statistical Inference*. Springer, 198.
- Minoru Kobayashi and Chris Schmandt, "Dynamic Soundscape: Mapping time to space for audio browsing," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1997, pp. 194–201.
- J. L. Mistler-Lachman, "Depth of comprehension and sentence memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 13, no. 1, pp. 98–106, 1974.
- D. Moffat and J. D. Reiss, "Perceptual Evaluation of Synthesized Sound Effects," *ACM Transactions on Applied Perception*, vol. 15, no. 2, pp. 1–19, 2018.
- A. R. Møller, *Hearing: Anatomy, Physiology, and Disorders of the Auditory System*. Plural Publishing, 2006.
- N. Moray, "Attention in dichotic listening: Affective cues and the influence of instructions," *Quarterly Journal of Experimental Psychology*, vol. 11, no. 1, pp. 56–60, 1959.
- A. T. Mullins, "Audiostreamer: Leveraging The Cocktail Party Effect for Efficient Listening," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.
- J. Mycroft, T. Stockman, and J. D. Reiss, "Visual Information Search in Digital Audio Workstations," in *140th Convention of the Audio Engineering Society*. Audio Engineering Society, 2016.
- T. Nakano, "PlaylistPlayer : An Interface Using Multiple Criteria to Change the Playback Order of a Music Playlist," in *21st International Conference on Intelligent User Interfaces*. ACM, 2016, pp. 186–190.

- NASA, "NASA-TLX," <https://humansystems.arc.nasa.gov/groups/TLX/>, 2018, [Online; accessed 22-Dec-2018].
- , "NASA Task Load Index Sheet," <https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf>, 2018, [Online; accessed 22-Dec-2018].
- T. Neil, *Designing web interfaces: Principles and patterns for rich interactions*. O'Reilly Media, Inc., 2009.
- M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.
- J. A. Obermeyer and L. A. Edmondsa, "Attentive reading with constrained summarization adapted to address written discourse in people with mild aphasia," *American Journal of Speech-Language Pathology*, vol. 27, no. 1S, pp. 392–405, 2018.
- P. Parente, "Clique: Perceptually based, task oriented auditory display for GUI applications," Ph.D. dissertation, The University of North Carolina at Chapel Hill, 2008.
- E. K. Park, K. M. Lee, and D. H. Shin, "Social Responses to Conversational TV VUI," *International Journal of Technology and Human Interaction*, vol. 11, no. 1, pp. 17–32, 2015.
- D. Patel, D. Ghosh, and S. Zhao, "Teach Me Fast: How to Optimize Online Lecture Video Speeding for Learning in Less Time?" in *Proceedings of the Sixth International Symposium of Chinese CHI*. ACM, 2018, pp. 160–163.
- N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh, "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 733–742.

- M. Paule-Ruiz, V. Álvarez-García, J. R. Pérez-Pérez, and M. Riestra-González, "Voice interactive learning: A framework and evaluation," in *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education*. ACM, 2013, pp. 34–39.
- J. Peissig and B. Kollmeier, "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," *The Journal of the Acoustical Society of America*, vol. 101, no. 3, pp. 1660–1670, 1997.
- S. Perugini, T. J. Anderson, and W. F. Moroney, "A study of out-of-turn interaction in menu-based, IVR, voicemail systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2007, p. 961.
- R. Pettersson, "Introduction to Message Design," *Journal of Visual Literacy*, vol. 31, no. 2, pp. 93–104, 2012.
- R. Pettersson, *Information Design: Message Design*. Institute for Infology, 2013.
- R. Plomp and A. M. Mimpen, "Speech-reception threshold for sentences as a function of age and noise level," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1333–1342, 1979.
- R. Plompp and A. M. Mimpen, "Effect of the Orientation of the Speaker's Head and the Azimuth of a Noise Source on the Speech-Reception Threshold for Sentences," *Acta Acustica united with Acustica*, vol. 48, no. 5, pp. 325–328, 1981.
- I. Pollack and J. M. Pickett, "Stereophonic Listening and Speech Intelligibility against Voice Babble," *The Journal of the Acoustical Society of America*, vol. 30, no. 2, pp. 131–133, 1958.
- S. Pollmann, M. Maertens, D. Y. Von Cramon, J. Lepsien, and K. Hugdahl, "Dichotic listening in patients with splenial and nonsplenial callosal lesions," *Neuropsychology*, vol. 16, no. 1, pp. 56–64, 2002.

- S. Poslad, *Ubiquitous Computing: Smart Devices, Environments and Interactions - Poslad - Wiley Online Library*. John Wiley & Sons, 2011.
- B. Qudah and N. J. Sarhan, "Efficient delivery of on-demand video streams to heterogeneous receivers," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 6, no. 3, pp. 1–25, 2010.
- N. R, "Two-sided confidence intervals for the single proportion: Comparison of seven methods," *Statistics in Medicine*, pp. 857–872, 1998.
- R. A. Rasch, "The perception of simultaneous notes such as in polyphonic music," *Acta Acustica united with Acustica*, vol. 40, no. 1, pp. 21–33, 1978.
- G. Rhodes, "Auditory attention and the representation of spatial information," *Perception & Psychophysics*, vol. 42, no. 1, pp. 1–14, 1987.
- M. Rivenez, C. J. Darwin, and A. Guillaume, "Processing unattended speech," *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4027–4040, 2006.
- M. Rivenez, A. Guillaume, L. Bourgeon, and C. J. Darwin, "Effect of voice characteristics on the attended and unattended processing of two concurrent messages," *European Journal of Cognitive Psychology*, vol. 20, no. 6, pp. 967–993, 2008.
- D. Ronzani, "The Battle of Concepts: Ubiquitous Computing, Pervasive Computing and Ambient Intelligence in Mass Media," *Ubiquitous Computing and Communication Journal*, vol. 4, no. 2, pp. 9–19, 2009.
- M. M. Rose and B. C. J. Moore, "Effects of frequency and level on auditory stream segregation," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, p. 1209, 2000.
- M. M. Rosee and B. C. J. Moore, "Effects of frequency and level on auditory stream segregation," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, p. 1209, 2000.

- P. Sanderson, "The multimodal world of medical monitoring displays," *Applied Ergonomics*, vol. 37, no. 4, pp. 501–512, 2006.
- D. Sato, S. Zhu, M. Kobayashi, H. Takagi, and C. Asakawa, "Sasayaki: augmented voice web browsing experience," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 2769–2778.
- C. Schmandt and A. Mullins, "AudioStreamer: Exploiting simultaneity for listening," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1995, pp. 218–219.
- C. Schmandt, "Audio hallway: a virtual acoustic environment for browsing," in *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*. ACM, 1998, pp. 163–170.
- B. A. Schneider, K. Pichora-Fuller, and M. Daneman, "Effects of Senescent Changes in Audition and Cognition on Spoken Language Comprehension," in *The aging auditory system*. Springer, 2010, pp. 167–210.
- D. Schnelle-Walka, "I tell you something," in *Proceedings of the 16th European Conference on Pattern Languages of Programs*. ACM, 2012, pp. 1–10.
- J. H. Schuett and B. N. Walker, "Measuring comprehension in sonification tasks that have multiple data streams," in *Proceedings of the 8th Audio Mostly Conference*. ACM, 2013, pp. 1–6.
- J. H. Schuett, R. J. Winton, J. M. Batterman, and B. N. Walker, "Auditory weather reports: demonstrating listener comprehension of five concurrent variables," in *Proceedings of the 9th Audio Mostly*. ACM, 2014, p. 17.
- B. Shinn-Cunningham, "Auditory precedence effect," *Encyclopedia of Computational Neuroscience*, pp. 252–253, 2015.

- B. Shneiderman, *Designing the user interface: strategies for effective human-computer interaction*. Pearson, 2010.
- E. Sikström and J. Berg, "Designing auditory display menu interfaces-cues for users current location in extensive menus," in *126th Convention of the Audio Engineering Society*. Audio Engineering Society, 2009.
- S. A. Simpson and M. Cooke, "Consonant identification in N-talker babble is a nonmonotonic function of N," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 2775–2778, 2005.
- J. Sodnik, G. Jakus, and S. Tomažic, "The use of spatialized speech in auditory interfaces for computer users who are visually impaired," *Journal of Visual Impairment & Blindness*, vol. 106, pp. 634–645, 2012.
- H. J. Song and K. Beilharz, "Aesthetic and auditory enhancements for multi-stream information sonification," in *Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts*. ACM, 2008, pp. 224–231.
- A. W. Stedmon, *Practical Speech User Interface Design*. CRC Press, Inc., 2015.
- A. Q. Summerfield and J. F. Culling, "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency," *The Journal of the Acoustical Society*, vol. 92, no. 4, p. 2317, 1992.
- D. Tilman, L. Jeffrey, and W. Bruce N, "Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons and speech," in *14th International Conference on Auditory Display (ICAD'06)*. Georgia Institute of Technology, 2008, pp. 1–6.
- Timothy D. Griffiths and Jason D. Warre, "What is an auditory object?" *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 887–892, 2004.

- J. A. Towers, "Enabling the Effective Application of Spatial Auditory Displays in Modern Flight Decks," Ph.D. dissertation, The University of Queensland, 2016.
- A. Treisman and R. Squire, "Listening to speech at two levels at once." *The Quarterly Journal of Experimental Psychology*, vol. 26, no. 1, pp. 82–97, 1974.
- A. M. Treisman, "Selective Attention in Man," *British Medical Bulletin*, vol. 20, no. 1, pp. 12–16, 1964.
- , "Strategies and models of selective attention." *Psychological review*, vol. 76, no. 3, p. 282, 1969.
- S. Truschin, M. Schermann, S. Goswami, and H. Krcmar, "Designing interfaces for multiple-goal environments," *ACM Transactions on Computer-Human Interaction*, vol. 21, no. 1, pp. 1–24, 2014.
- L. P. A. S. van Noorden, *Temporal coherence in the perception of tone sequences*. Institute for Perceptual Research Eindhoven, The Netherlands, 1975.
- , "Minimum differences of level and frequency for perceptual fission of tone sequences ABAB," *The Journal of the Acoustical Society of America*, vol. 61, no. 4, pp. 1041–1045, 1977.
- Y. Vazquez Alvarez and S. A. Brewster, "Designing spatial audio interfaces to support multiple audio streams," in *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, 2010, pp. 253–256.
- Y. Vazquez-Alvarez, M. P. Aylett, S. A. Brewster, R. von Jungefeld, and A. Virolainen, "Multilevel auditory displays for mobile eyes-free location-based interaction," in *Proceedings of the Extended Abstracts of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 1567–1572.

- Y. Vazquez-Alvarez, M. P. Aylett, S. A. Brewster, R. V. Jungenfeld, and A. Viro-lainen, "Designing interactions with multilevel auditory displays in mobile audio-augmented reality," *ACM Transactions on Computer-Human Interaction*, vol. 23, no. 1, pp. 1–30, 2015.
- S. R. Walter, M. Z. Raban, W. T. Dunsmuir, H. E. Douglas, and J. I. Westbrook, "Emergency doctors' strategies to manage competing workload demands in an interruptive environment: An observational workflow time study," *Applied Ergonomics*, vol. 58, pp. 454–460, 2017.
- C. Wasylyshyn, B. McClimens, and D. Brock, "Comprehension of speech presented at synthetically accelerated rates: Evaluating training and practice effects," in *Proceedings of the 16th International Conference on Auditory Display (ICAD)*. Georgia Institute of Technology, 2010, pp. 133–136.
- Web-AudioCheck, "AudioCheck," <http://www.audiocheck.net>, 2016, [Online; accessed 15-July-2015].
- C. L. Webber and N. Marwan, *Recurrence Quantification Analysis Theory and Best Practices*. Springer, 2015.
- D. Weber and M. Phillips, "An Architecture for Multi-View Information Overlays," in *Proceedings of the 2004 Australasian Symposium on Information Visualisation*. Australian Computer Society, Inc., 2004, pp. 9–15.
- M. Weiser, "The Computer for 21 st Century," *Scientific American*, vol. 265, no. 3, pp. 94–105, 1991.
- R. J. Welland, R. Lubinski, and D. J. Higginbotham, "Discourse Comprehension Test Performance of Elders With Dementia of the Alzheimer Type," *Journal of Speech Language and Hearing Research*, vol. 45, no. 6, p. 1175, 2002.

- S. Werner, C. Hauck, N. Roome, C. Hoover, and D. Choates, "Can VoiceScapes assist in menu navigation?" in *Proceedings of the Human Factors and Ergonomics Society*, vol. 2015, no. 1, 2015, pp. 1095–1099.
- R. Westerhausen and K. Hugdahl, "The corpus callosum in dichotic listening studies of hemispheric asymmetry: A review of clinical and experimental evidence," *Neuroscience and Biobehavioral Reviews*, vol. 32, no. 5, pp. 1044–1054, 2008.
- R. Westerhausen and K. Kompus, "How to get a left-ear advantage: A technical review of assessing brain asymmetry with dichotic listening," *Scandinavian journal of psychology*, vol. 59, no. 1, pp. 66–73, 2018.
- J. White and M. Duggirala, "Speech-interface prompt design," in *Proceedings of the 7th International Conference on Information and Communication Technologies and Development*. ACM, 2015, pp. 1–4.
- S. M. Williams, *Perceptual Principles in Sound Grouping*. Addison-Wesley, 1994.
- E. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, pp. 209–212, 1927.
- W. World Wide Web Consortium, "W3C Speech Interface Framework," <https://www.w3.org/TR/voice-intro/>, [online, accessed on 13 Mar 2015].
- T. Wu, W. Dou, F. Wu, S. Tang, C. Hu, and J. Chen, "A Deployment Optimization Scheme Over Multimedia Big Data for Large-Scale Media Streaming Application," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 5s, pp. 1–23, 2016.
- J. Xia, N. Nooraei, S. Kalluri, and B. Edwards, "Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 1888–1898, 2015.

- C. Xu, N. C. Maddage, X. Shao, and Q. Tian, "Content-adaptive digital music watermarking based on music structure analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 1–es, 2007.
- B. S. Yandell, *Practical Data Analysis for Designed Experiments*. Chapman Hall, 1997.
- W. A. Yost, *Fundamentals of hearing: An introduction*. Academic Press, 1994.
- , "The cocktail party problem: Forty years later," *Binaural and Spatial Hearing in Real and Virtual Environments*, pp. 329–347, 1997.
- W. A. Yost, R. H. Dye, and S. Sheft, "A simulated "cocktail party" with up to three sound sources," *Perception and Psychophysics*, vol. 58, no. 7, pp. 1026–1036, 1996.
- J. C. Yu, T. Y. Chang, and C. T. Yang, "Individual differences in working memory capacity and workload capacity," *Frontiers in Psychology*, vol. 5, pp. 699–703, 2014.
- M. Yusro, K. M. Hou, E. Pissaloux, K. Ramli, D. Sudiana, L. Z. Zhang, and H. L. Shi, "Concept and Design of SEES (Smart Environment Explorer Stick) for Visually Impaired Person Mobility Assistance," *Advances in Intelligent Systems and Computing*, vol. 300, pp. 245–259, 2014.
- N. Zhang, H. Huang, B. Su, J. Zhao, and B. Zhang, "Information dissemination analysis of different media towards the application for disaster pre-warning," *PLOS ONE*, vol. 9, no. 5, pp. 45–80, 2014.
- R. Zimmermann and K. Liang, "Spatialized audio streaming for networked virtual environments," in *Proceeding of the 16th ACM International Conference on Multimedia*. ACM, 2008, pp. 299–308.

Appendix A

A Letter to the Institute Requesting the Participation of Visually Challenged Persons in Study II

**National Training Centre For Special Persons,
G-9/2, Islamabad**

Ref. No. DCS(C-5) / 2016 -

Dated: April. 28, 2016

Subject: Exploring possibilities to provide multiple voice-based information to the visually challenged persons

Dear Sir / Madam,

The Department of Computer Sciences at Quaid-i-Azam University Islamabad has been exploring the possibilities to aid the visually challenged person in reaching information quickly using their auditory capabilities. For this, one of our doctoral students Muhammad Abu ul Fazal is carrying out a research study that involve a series of experiments with visually challenged persons focusing on how they interact with information.

We would be highly obliged if you could allow our scholar to visit your organization for the purpose and allow to carry out experiments with your pupils occasionally. We assure you about confidentiality of the information thus gathered and would keep you updated about progress made in terms of research publications and / or sharing of tools thus developed in our Lab.

I thank you in anticipation for your cooperation in this regard.

Best regards,

Research Supervisor

Dr. Muhammad Shuaib Karim
Assistant Professor,
Tel (Sec.): 051-90642057, (Off.): 051-90642055
Email: skarim@qau.edu.pk

Appendix B

Vinformize-based System Interface

Multiple Information Communication in Voice Based Interaction

This is a research project to investigate the design possibility of communicating multiple voice-based information simultaneously.

Profile (jk@jk.com) | Logout

Play **Fast Forward / Back** **Volume**

Pan **Playback-rate**

Left Ear Mid Right Ear 0.5x 1 2x

Documentary

- Hawking 'transformed our view of the universe' - BBC News
- Bitcoin explained: How do cryptocurrencies
- Davos explained: In 90 seconds

Play **Fast Forward / Back** **Volume**

Pan **Playback-rate**

Left Ear Mid Right Ear 0.5x 1 2x

News

- 2018 April 01 BBC One minute World News
- 2018 march 30 BBC One minute World News
- 2018 march 14 BBC One minute World News
- 2018 march 05 BBC One Minute World News

Play **Fast Forward / Back** **Volume**

Pan **Playback-rate**

Left Ear Mid Right Ear 0.5x 1 2x

Interview

- Mark Zuckerberg in 2009: Facebook
- Sachin Tendulkar Full Interview
- The Briton taking Bollywood by storm
- Malala: I am so happy to be home

Play **Fast Forward / Back** **Volume**

Pan **Playback-rate**

Left Ear Mid Right Ear 0.5x 1 2x

Sports Commentary

- Foot ball 1
- Foot ball 2
- Foot ball 3
- Foot ball 4

Play **Fast Forward / Back** **Volume**

Pan **Playback-rate**

Left Ear Mid Right Ear 0.5x 1 2x

Songs

- BGM: Displacement
- Alexandra Stan - Mr Saxobeat
- BGM: Gandhi
- Rihanna - Whats My Name ft Drake

Appendix C

Ethics Application for Study III

UTS Creativity and Cognition Studios

3-page Ethics Approval Application

From: *Muhammad Abu ul Fazal, Dr. Andrew Johnston, Dr. Sam Ferguson*

Project Number: 2016-6, HREC 2013000135

1. Title

Listening to Multiple Voice-based Information Streams

2. Aims

Studying users' abilities of Listening to Multiple Voice-based Information Streams to get data that addresses the following questions:

- Are users capable of listening and comprehending multiple voice-based information streams played simultaneously?
- How well do users comprehend multiple information streams presented using various techniques?
- Which information presentation helps users more to understand and comprehend multiple streams efficiently?
- Is this method of communication helpful to fulfill users' information needs?

3. Methodology

The following methods will be applied:

- A number of audio stimuli will be developed to reflect various information presentation techniques. These will include more than one voice talking simultaneously.
- A web-based application will be developed using PHP, MySQL, HTML CSS, Bootstrap and Javascript to play the stimuli and record user's response to questionnaires in database
- The experiment would be conducted in the Creativity & Cognition Studios where users will listen to the stimuli using standardized equipment (headphones and PC).

4. Significance

The post analysis of the users' responses would help to establish whether providing the multiple voice-based information simultaneously to the user is effective/understandable or not. The analysis will identify which type of information presentation, if any, is most suitable for the users to listen and understand the multiple information simultaneously.

5. Number of participants and justification of numbers

60

Since our experiment would investigate the possibility of communicating multiple voice-based information streams simultaneously to the typical users, therefore, a minimum of 60 number of motivated users would be required to participate in the experiment.

6. Selection/exclusion criteria

The users would mostly be from the Faculty of IT, UTS. Since all of the information will be in English native English speakers with no significant hearing impairment will be given preference to participate in the experiment to avoid language issues contributing to poor comprehension.

7. Children under 18 years of age will participate in the evaluation.

NO

8. Procedures

Users will be briefed about the system and given instructions on how to use it. The system will be designed to be user-friendly so that users should not find any difficulty in using it.

- Users will provide their profile information to the system that includes Name: Email: Age: Gender: Primary Language: Highest Qualification (eg. high school certificate, Bachelor degree, etc): Profession: Hearing Impairment (if any): Visual Impairment (if any):
- Each user will randomly be assigned with 12 out of 216 stimuli
- Each user will listen to stimuli on one of the three MAC computers and three DT 770 Pro headphones.
- After hearing each stimulus, a questionnaire will be answered by the user
- All the answers will be stored in a database for post analysis

9. Time commitment for participants

The average length of the stimulus is 70 seconds. That means a user would spend around 15 minutes to listen to all the stimuli and can take up to 45 minutes including breaks to answer the yes/no response questions.

Accumulatively, approximately 60 minutes will be required from a user to participate in the experiment.

10. Location of research

The experiment will be set up in the Creativity and Cognition Studios (CCS).

11. Consent procedures

Signed consent sheet (see attached)

12. Additional Risks (additional to those noted in the CCS Generic Approval)

none

13. Strategies to cope with risks mentioned in 12.

none

14. Funding source(s) & potential conflicts of interest

No conflicts of interest.

15. Strategies to cope with any conflicts of interest identified in 14.

Not applicable.

15. Other issues

No other issues perceived as being problematic.

*Number obtained from CCS Ethics Administrator prior to completing form.

UTS: IT: CREATIVITY & COGNITION STUDIOS

AUDIENCE STUDIESUTS HREC REF 2013000135

Listening to Multiple Voice-Based Information Streams, CCS Internal Application number 2016-6

GENERAL INFORMATION

WHO IS DOING THE RESEARCH?

My name is Muhammad Abu ul Fazal and I am a student at UTS. My supervisors are Dr. Andrew Johnston & Dr. Sam Ferguson.

WHAT IS THIS RESEARCH ABOUT?

This research is to estimate the comprehensibility of multiple voice-based audio streams played simultaneously.

IF I SAY YES, WHAT WILL IT INVOLVE?

I will ask you to

- give your basic profile information
- listen 12 audio stimuli on a Mac computer using a headphone
- answer a questionnaire containing 16 yes/no questions after each stimuli

ARE THERE ANY RISKS?

There are minimal risks involved in this research. The volume of the audio streams played will not exceed normal conversational levels.

WHY HAVE I BEEN ASKED?

Because you

- are a native English speaker; or,
- can hear and understand English language without any difficulty; and
- do not have any significant hearing impairment,

DO I HAVE TO SAY YES?

You don't have to say yes.

WHAT WILL HAPPEN IF I SAY NO?

Nothing. I will thank you for your time so far and won't contact you about this research again.

IF I SAY YES, CAN I CHANGE MY MIND LATER?

You can change your mind at any time and you don't have to say why. I will thank you for your time so far and won't contact you about this research again.

WHAT IF I HAVE CONCERNS OR A COMPLAINT?

If you have concerns about the research that you think I or my supervisor can help you with, please feel free to contact us on Muhammad.AbuUlFazal@student.uts.edu.au , andrew.johnston@uts.edu.au , Samuel.Ferguson@uts.edu.au

If you would like to talk to someone who is not connected with the research, you may contact the Research Ethics Officer via Research.Ethics@uts.edu.au, and quote this number: 2013000135.

This study has been approved by the University of Technology, Sydney, Human Research Ethics Committee.

If you have any complaints or reservations about any aspect of your participation in this research that you cannot resolve with the researcher, you may contact the UTS Ethics Committee through the Research Ethics Officer at UTS Broadway, Building 1, Level 14; or 9514 9772; or Research.Ethics@uts.edu.au. Please quote the UTS HREC reference number.

Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

UTS: CREATIVITY & COGNITION STUDIOS

CONSENT FORMUTS HREC REF NO 2013000135

Listening to Multiple Voice-based Information Streams, CCS Internal Application number 2016-6

I _____ (participant's name) agree to participate in the research project *Listening To Multiple Voice-based Information Streams* (HREC 2013000135 project number 2016-6) being conducted by *Muhammad Abu ul Fazal, Dr. Andrew Johnston and Dr. Sam Ferguson* of the Creativity and Cognition Studios at the University of Technology, Sydney.

I understand that the purpose of this study is to share my experience of Listening to Multiple Voice-based Audio Streams simultaneously.

I understand that I have been asked to participate in this research because I

- am a native English speaker
- can hear and understand English language without any difficulty
- do not have any significant hearing impairment,

I understand that my participation in this research will include

- giving my basic profile information
- listening to 12 audio stimuli on a Mac computer using headphones
- answering a questionnaire containing 16 yes/no questions after each stimuli

The system will store my profile information and answers to questions for post analysis.

I am aware that I can contact Muhammad Abu ul Fazal, [Muhammad.AbuUIFazal@student.uts.edu.au](mailto:Mohammad.AbuUIFazal@student.uts.edu.au) if I have any concerns about the research. I also understand that I am free to withdraw my participation from this research project at any time I wish, without consequences, and without giving a reason.

I agree that the research data gathered from this project may be published in a form that does not identify me in any way.

Signed by _____ / ____ / ____

Witnessed by _____ / ____ / ____

NOTE:

This study has been approved by the University of Technology, Sydney, Human Research Ethics Committee.

If you have any complaints or reservations about any aspect of your participation in this research that you cannot resolve with the researcher, you may contact the UTS Ethics Committee through the Research Ethics Officer at UTS Broadway, Building 1, Level 14; or 9514 9772; or Research.Ethics@uts.edu.au. Please quote the UTS HREC reference number.

Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

UTS Creativity and Cognition Studios

3-page Ethics Approval Application

From: *Muhammad Abu ul Fazal, Dr. Andrew Johnston, Dr. Sam Ferguson*

Project Number: 2016-6, HREC 2013000135

1. Title

Listening to Multiple Voice-based Information Streams

2. Aims

Studying users' abilities of Listening to Multiple Voice-based Information Streams to get data that addresses the following questions:

- Are users capable of listening and comprehending multiple voice-based information streams played simultaneously?
- How well do users comprehend multiple information streams presented using various techniques?
- Which information presentation helps users more to understand and comprehend multiple streams efficiently?
- Is this method of communication helpful to fulfill users' information needs?

3. Methodology

The following methods will be applied:

- A number of audio stimuli will be developed to reflect various information presentation techniques. These will include more than one voice talking simultaneously.
- A web-based application will be developed using PHP, MySQL, HTML CSS, Bootstrap and Javascript to play the stimuli and record user's response to questionnaires in database
- The experiment would be conducted in the Creativity & Cognition Studios where users will listen to the stimuli using standardized equipment (headphones and PC).

4. Significance

The post analysis of the users' responses would help to establish whether providing the multiple voice-based information simultaneously to the user is effective/understandable or not. The analysis will identify which type of information presentation, if any, is most suitable for the users to listen and understand the multiple information simultaneously.

5. Number of participants and justification of numbers

60

Since our experiment would investigate the possibility of communicating multiple voice-based information streams simultaneously to the typical users, therefore, a minimum of 60 number of motivated users would be required to participate in the experiment.

6. Selection/exclusion criteria

The users would mostly be from the Faculty of IT, UTS. Since all of the information will be in English native English speakers with no significant hearing impairment will be given preference to participate in the experiment to avoid language issues contributing to poor comprehension.

7. Children under 18 years of age will participate in the evaluation.

NO

8. Procedures

Users will be briefed about the system and given instructions on how to use it. The system will be designed to be user-friendly so that users should not find any difficulty in using it.

- Users will provide their profile information to the system that includes Name: Email: Age: Gender: Primary Language: Highest Qualification (eg. high school certificate, Bachelor degree, etc): Profession: Hearing Impairment (if any): Visual Impairment (if any):
- Each user will randomly be assigned with 12 out of 216 stimuli
- Each user will listen to stimuli on one of the three MAC computers and three DT 770 Pro headphones.
- After hearing each stimulus, a questionnaire will be answered by the user
- All the answers will be stored in a database for post analysis

9. Time commitment for participants

The average length of the stimulus is 70 seconds. That means a user would spend around 15 minutes to listen to all the stimuli and can take up to 45 minutes including breaks to answer the yes/no response questions.

Accumulatively, approximately 60 minutes will be required from a user to participate in the experiment.

10. Location of research

The experiment will be set up in the Creativity and Cognition Studios (CCS).

11. Consent procedures

Signed consent sheet (see attached)

12. Additional Risks (additional to those noted in the CCS Generic Approval)

none

13. Strategies to cope with risks mentioned in 12.

none

14. Funding source(s) & potential conflicts of interest

No conflicts of interest.

15. Strategies to cope with any conflicts of interest identified in 14.

Not applicable.

15. Other issues

No other issues perceived as being problematic.

*Number obtained from CCS Ethics Administrator prior to completing form.

Appendix D

Interface with Selected Playable Audio Files URLs, and Questionnaire for Study III

All the web URLs mentioned in the interface are playable stimuli (open in web browser)
representing relevant designs.



Profile

You need to Login / Register first.

 Select Age 

 Select Gender 

 Select Primary Language 

 Qualification 

 Profession / Discipline 

 Select Country 

By clicking the check box I agree to participate in the this research experiment "Listening To Multiple Voice-based Information Streams (HREC 2013000135 project number 2016-6)". To view consent form [click here](#) (PDF file opens in another Tab).

Instructions



Instructions

1) You are required to listen to all 14 players given on this page. Each player plays multiple files in female and male voices in a different presentation. The presentation type is stated in the title given on top of each player.

2) Start listening to content from Player 1. After that click on the link 'Show Questions for this Player: 1' given under the player-1. It would open 16 yes/no questions that you are required to answer.

3) Repeat this process to remaining 13 players in incremental order. In the end submit the response by pressing the 'Submit Response' button.

4) The content loading may take some depending on the Internet speed. After loading of content a play sign will appear on each player. 5) Please do not listen the players more than once.

I have read the Instructions.

Player: 1, **Type:** Both the voices are continuous, **Presentation:** Female & Male Voices to the both ears



<https://bit.ly/2KaY5pc>

Show Questions for this Player: 1

Car Stalled,

- 1. Was Jim an **insurance salesman**? DTS
 Yes | No | Don't Know
- 2. Did Jim's car stop running in the **afternoon**? DTI
 Yes | No | Don't Know
- 3. Was Jim driving through the **city**? MIS
 Yes | No | Don't Know
- 4. Did Jim eventually **start walking to a gas station**? MS
 Yes | No | Don't Know
- 5. Was Jim driving a **fairly new car**? DTI
 Yes | No | Don't Know
- 6. Was Jim's car **out of gas**? MI
 Yes | No | Don't Know
- 7. Did Jim stop to talk to an **old** man? DTS
 Yes | No | Don't Know
- 8. Was there a gas station **close to where Jims car stopped running**? MI
 Yes | No | Don't Know

Husban Wife,

- 113. Did this story happen in the **summer**? DTI
 Yes | No | Don't Know
- 114. Was Betty **reading a book**? MIS
 Yes | No | Don't Know
- 115. Was Joe **watching TV**? DTS
 Yes | No | Don't Know
- 116. Did Joe and Betty **have children**? DTI
 Yes | No | Don't Know
- 117. Did Joe ask his wife to bring him **something to eat**? MIS
 Yes | No | Don't Know
- 118. Did Betty make Joe a **tuna sandwich**? DTS
 Yes | No | Don't Know
- 119. Did Joe eat the **sandwich and cookies**? MI
 Yes | No | Don't Know
- 120. Did Betty **usually make Joe get things for himself**? MI
 Yes | No | Don't Know

Player: 2, **Type:** Both the voices are continuous, **Presentation:** Female voice to the both ears and male voice to the Right Ear only



<https://bit.ly/2znprnV>

Show Questions for this Player: 2

SandWhich,

- 33. Did Sam **drive his car** downtown? DTS
 Yes | No | Don't Know
- 34. Was Sam a **young** man? DTI
 Yes | No | Don't Know
- 35. Did Sam pick up a **watch** at the jewelry store? DTS
 Yes | No | Don't Know
- 36. Did Sam look for a place to have **lunch**? MS
 Yes | No | Don't Know
- 37. Did Sam have a **college degree**? DTI
 Yes | No | Don't Know
- 38. Did Sam **go into restaurant with a sign in the window**? MI
 Yes | No | Don't Know
- 39. Did Sam order a **ham** sandwich? MS
 Yes | No | Don't Know
- 40. Did Sam **get the kind of sandwich he ordered**? MI
 Yes | No | Don't Know

New Business,

- 105. Were Fred and Ben **brothers**? DTS
 Yes | No | Don't Know
- 106. Did Fred and Ben start a **lawn care business**? MS
 Yes | No | Don't Know
- 107. Did Mrs. Foster call the man on **Sunday Morning**? DTI
 Yes | No | Don't Know
- 108. Was Mrs. Foster **their first Customer**? MS
 Yes | No | Don't Know
- 109. Was Mrs. Foster **in a hurry** to have her house painted? MI
 Yes | No | Don't Know
- 110. Did Fred and Ben promise to have the house painted **by Thursday**? DTS
 Yes | No | Don't Know
- 111. Were Fred and Ben painting a **two-story house**? DTI
 Yes | No | Don't Know
- 112. Were Fred and Ben painting **the right house**? MI
 Yes | No | Don't Know

Player: 3, **Type:** Both the voices are continuous, **Presentation:** Female voice to the left ear only and male voice to the Right Ear only



<http://bit.ly/2S1FJtr>

Show Questions for this Player: 3

Lost Book Library,

- 17. Did Henry go to **a bookstore**? MS
 Yes | No | Don't Know
- 18. Did this story happen on **a sunny day**? DTI
 Yes | No | Don't Know
- 19. Did the librarian have **gray hair**? DTS
 Yes | No | Don't Know
- 20. Was Henry's book **overdue**? MI
 Yes | No | Don't Know
- 21. Did Henry **lose the book**? MS
 Yes | No | Don't Know
- 22. Was Henry **on vacation** when he lost the book? DTI
 Yes | No | Don't Know
- 23. Did Henry leave the book **in Florida**? DTS
 Yes | No | Don't Know
- 24. Did Henry **find the book**? MI
 Yes | No | Don't Know

Loan,

- 97. Was Neil a **high school** student? MI
 Yes | No | Don't Know
- 98. Did Neil's parents **live nearby**? DTI
 Yes | No | Don't Know
- 99. Did Neil go to the bank **to get a loan**? MS
 Yes | No | Don't Know
- 100. Did Neil need the money **to start a new business**? MS
 Yes | No | Don't Know
- 101. Did Neil own a **car**? DTS
 Yes | No | Don't Know
- 102. Did Neil go to the bank **in the morning**? DTI
 Yes | No | Don't Know
- 103. Did Neil tell the woman that he had a **cheese sandwich** for lunch? DTS
 Yes | No | Don't Know
- 104. Did Neil **get the loan**? MI
 Yes | No | Don't Know

Player: 4, **Type:** Female voice is continuous and male voice is in chunks, **Presentation:** Female & Male Voices to the both ears



<http://bit.ly/2FrGcDR>

Show Questions for this Player: 4

Fire,

- 41. Did Mrs. Wilson **live alone** on a farm? MS
 Yes | No | Don't Know
- 42. Did Mrs. Wilson live in **Kansas**? DTS
 Yes | No | Don't Know
- 43. Did Mrs. Wilson's son want her to **move in with him**? DTS
 Yes | No | Don't Know
- 44. Did **lighting strike the barn**? MI
 Yes | No | Don't Know
- 45. Was Mrs. Wilson's phone **in her bedroom**? DTI
 Yes | No | Don't Know
- 46. Did the fire **go out by itself**? MI
 Yes | No | Don't Know
- 47. Were there **cows in the Barn**? DTI
 Yes | No | Don't Know
- 48. Did **Mrs. Wilson** call the fire department? MS
 Yes | No | Don't Know

Travel Airport,

- 121. Was Don waiting at **a doctor's office**? MS
 Yes | No | Don't Know
- 122. Was Don flying to **his parent's home**? DTI
 Yes | No | Don't Know
- 123. Was Don's **sister** going to be at his parent's home? DTS
 Yes | No | Don't Know
- 124. Were weather conditions **good**? DTI
 Yes | No | Don't Know
- 125. Were **some people told to wait for the next plane**? MS
 Yes | No | Don't Know
- 126. Was Don told that **he would have to wait for the next plane**? MI
 Yes | No | Don't Know
- 127. Was Don flying to **Denver**? DTS
 Yes | No | Don't Know
- 128. Did the woman **believe Don's story**? MI
 Yes | No | Don't Know

Player: 5, **Type:** Female voice is continuous and male voice is in chunks, **Presentation:** Female voice to the both ears and male voice to the Right Ear only



<http://bit.ly/2FrGdYr>

Show Questions for this Player: 5

Chalan Ticket,

- 9. Did this story happen in the **summer**? D11
 Yes | No | Don't Know
- 10. Was Harry driving a **truck**? DTS
 Yes | No | Don't Know
- 11. Was Harry going to the **cleaners**? MIS
 Yes | No | Don't Know
- 12. Did the policeman ask to see Harry's car **registration**? MIS
 Yes | No | Don't Know
- 13. Did Harry have **a lot of cards** in his wallet? D11
 Yes | No | Don't Know
- 14. Had Harry's driver's license been expired for **a month**? DTS
 Yes | No | Don't Know
- 15. Did the policeman tell Harry that **he had been speeding**? M11
 Yes | No | Don't Know
- 16. Did Harry **make it to the cleaners before they closed**? M11
 Yes | No | Don't Know

Garage Sale,

- 129. Did several women **have a party**? MIS
 Yes | No | Don't Know
- 130. Were there **a large number of things** at the garage sale? M11
 Yes | No | Don't Know
- 131. Did the women put up a sign **at a shopping center**? DTS
 Yes | No | Don't Know
- 132. Was it **cold** the day of the garage sale? D11
 Yes | No | Don't Know
- 133. Was the man driving a **car**? DTS
 Yes | No | Don't Know
- 134. Was the mattress **in terrible condition**? MIS
 Yes | No | Don't Know
- 135. Was the man **married**? D11
 Yes | No | Don't Know
- 136. Was the man **fond of his father-in-law**? M11
 Yes | No | Don't Know

Player: 6, **Type:** Female voice is continuous and male voice is in chunks, **Presentation:** Female voice to the left ear only and male voice to the Right Ear only



<http://bit.ly/2FpNzeX>

Show Questions for this Player: 6

Stunt,

- 25. Was George **a bookkeeper**? DTS
 Yes | No | Don't Know
- 26. Was George planning to **walk between two buildings** on a tightrope? MIS
 Yes | No | Don't Know
- 27. Had George **tried this stunt before**? MI
 Yes | No | Don't Know
- 28. Was George **single**? DTI
 Yes | No | Don't Know
- 29. Did George's wife **try to convince him not to try this stunt**? MIS
 Yes | No | Don't Know
- 30. Did **about a hundred people** gather to watch George? DTS
 Yes | No | Don't Know
- 31. Did George start to cross the falls **from the American side**? DTI
 Yes | No | Don't Know
- 32. Did George **make it across the falls**? MI
 Yes | No | Don't Know

Baseball,

- 137. Did George go to **a baseball game**? MSI
 Yes | No | Don't Know
- 138. Was it **an evening game**? DTI
 Yes | No | Don't Know
- 139. Was George's baseball glove **new**? DTSI
 Yes | No | Don't Know
- 140. Did George **try to impress the people around him**? MI
 Yes | No | Don't Know
- 141. Did the batter hit **a foul ball**? DTSI
 Yes | No | Don't Know
- 142. Was George sitting **in the front row**? DTI
 Yes | No | Don't Know
- 143. Did George **catch the ball**? MI
 Yes | No | Don't Know
- 144. Was the man **a baseball scout**? MSI
 Yes | No | Don't Know

Player: 7, **Type:** Both the voices are continuous, **Presentation:** Female & Male Voices to the both ears



<http://bit.ly/2TiCMWF>

Show Questions for this Player: 7

Researcher Experience,

- 57. Did speaker talk about the **doctors and medical facilities**? MIS
 Yes | No | Don't Know
- 58. Did speaker speak of the **emergency ward in a hospital**? MIS
 Yes | No | Don't Know
- 59. According to the speaker, are the researchers from different countries **not helpful** for research activities? MI
 Yes | No | Don't Know
- 60. Did speaker to set up the research program in **one** country only? MI
 Yes | No | Don't Know
- 61. Did speaker contacted fellows by **phone call**? DTS
 Yes | No | Don't Know
- 62. Did speaker contact **undergraduates** for assistance? DTS
 Yes | No | Don't Know
- 63. Is speaker a **researcher**? DTI
 Yes | No | Don't Know
- 64. Was it a **pleasant experience** of the researcher when he contacted others for help? DTI
 Yes | No | Don't Know

Agriculture Park,

- 145. Did the speaker describe an **Agriculture Park**? MIS
 Yes | No | Don't Know
- 146. Did the speaker specifically talk about how to get a bumper **cotton crop**? MIS
 Yes | No | Don't Know
- 147. Is the place **suitable to visit by Agriculturists or Scientists**? MI
 Yes | No | Don't Know
- 148. Is the place organised into **different sections**? MI
 Yes | No | Don't Know
- 149. Has the place opened a **month** ago? DTS
 Yes | No | Don't Know
- 150. Does the place has a **single** variety of Animals? DTS
 Yes | No | Don't Know
- 151. Is the giant wall plan situated in **Reception block**? DTI
 Yes | No | Don't Know
- 152. Is the rare breed section **far away** from the Reception block? DTI
 Yes | No | Don't Know

Player: 8, **Type:** Both the voices are continuous, **Presentation:** Female voice to the both ears and male voice to the Right Ear only



<http://bit.ly/2Ftglew>

Show Questions for this Player: 8

Company Representative,

- 81. Did speaker talk about a **company**? MIS
 Yes | No | Don't Know
- 82. Was it a **construction** Company? MIS
 Yes | No | Don't Know
- 83. Does company offer **cross-countries** facilities? MII
 Yes | No | Don't Know
- 84. Does company has Yes experience in providing services? MII
 Yes | No | Don't Know
- 85. Is the company **at least a decade old**? DTS
 Yes | No | Don't Know
- 86. Does company has majority of sites in **France**? DTS
 Yes | No | Don't Know
- 87. Does Company offer facilities in **Australia**? DTI
 Yes | No | Don't Know
- 88. Does company provide services in more than **600** places? DTI
 Yes | No | Don't Know

Radio Art Center,

- 153. Is the speaker a **radio program host** MS?
 Yes | No | Don't Know
- 154. Did the speaker talk about **vaccinations** to the newborn? MS
 Yes | No | Don't Know
- 155. Does the program cover about the **events that are on offer in the coming week**? MII
 Yes | No | Don't Know
- 156. Is the center a **multipurpose building**? MII
 Yes | No | Don't Know
- 157. Is the name of the center "**The National Arts Centre**" DTS?
 Yes | No | Don't Know
- 158. Does center has a **public library**? DTS
 Yes | No | Don't Know
- 159. Is there a **sports stadium part of the center** to hold a sports event DTI?
 Yes | No | Don't Know
- 160. Can the center be visited to **watch a movie** DTI?
 Yes | No | Don't Know

Player: 9, **Type:** Both the voices are continuous, **Presentation:** Female voice to the left ear only and male voice to the Right Ear only



<http://bit.ly/2zYWf5P>

Show Questions for this Player: 9

Business System Lecture,

- 73. Did speaker talk about **business**? MIS
 Yes | No | Don't Know
- 74. Did speaker talk about how to get **admission in Faculty of Business Administration**? MS
 Yes | No | Don't Know
- 75. Is the speaker expert in **Security Services**? MI
 Yes | No | Don't Know
- 76. According to the speaker, is following the best practices in starting new businesses **always** remain successful? MI
 Yes | No | Don't Know
- 77. Did speaker gave an example of **managing bank branch**? DTS
 Yes | No | Don't Know
- 78. According to the speaker, does getting things implement for 2nd time **always remain easier** than the first time? DTS
 Yes | No | Don't Know
- 79. Was it a **morning** time? DTI
 Yes | No | Don't Know
- 80. Was it the **first** lecture of speaker on business? DTI
 Yes | No | Don't Know

HoneyBee,

- 161. Did speaker talk about **Quarantine Service**? MIS
 Yes | No | Don't Know
- 162. Did speaker describe his role w.r.t **computer anti-virus**? MIS
 Yes | No | Don't Know
- 163. Does the country **allow** any food from outside? MI
 Yes | No | Don't Know
- 164. Does the country **welcome** any insect pests in its territory? MI
 Yes | No | Don't Know
- 165. Was it a role of the speaker to **find honeybee and eradicate it**? DTS
 Yes | No | Don't Know
- 166. Had the speaker discovered the honeybees **previously** in Australia.? DTS
 Yes | No | Don't Know
- 167. Was it an **evening time** when speaker gave his talk? DTI
 Yes | No | Don't Know
- 168. Is the speaker a **researcher**? DTI
 Yes | No | Don't Know

Player: 10, **Type:** Female voice is continuous and male voice is in chunks, **Presentation:** Female & Male Voices to the both ears



<http://bit.ly/2PBldTG>

Show Questions for this Player: 10

Library Visit,

- 89. Did the speaker discuss about the **library**? MS
 Yes | No | Don't Know
- 90. Did speaker tell people how to **get a book issued from the library**? MS
 Yes | No | Don't Know
- 91. Were there the **new** visitors who visited the place? MII
 Yes | No | Don't Know
- 92. Does the place has **more than one room**? MII
 Yes | No | Don't Know
- 93. Was the speaker desk located on the **left** side of the visitors? DTS
 Yes | No | Don't Know
- 94. Does the place has reference books? DTS
 Yes | No | Don't Know
- 95. Did the speaker **used map/plan** to explain the property/place? DTI
 Yes | No | Don't Know
- 96. Are the speaker office and the reference books room located **opposite** to each other? DTI
 Yes | No | Don't Know

Study Scholarship,

- 185. Did speaker talk about his **grant**? MIS
 Yes | No | Don't Know
- 186. Did speaker no talk to his **parents**? MIS
 Yes | No | Don't Know
- 187. Had the speaker applied for the grant a **long ago**? MII
 Yes | No | Don't Know
- 188. Did speaker given an impression that Government provides **funding to all**? MII
 Yes | No | Don't Know
- 189. Did speaker request his **boss** for sponsorship? DTS
 Yes | No | Don't Know
- 190. Did speaker's parents provide him **support**? DTS
 Yes | No | Don't Know
- 191. Did speaker worked as an **employee**? DTI
 Yes | No | Don't Know
- 192. Did speaker contact the **local council**? DTI
 Yes | No | Don't Know

Player: 11, **Type:** Female voice is continuous and male voice is in chunks, **Presentation:** Female voice to the both ears and male voice to the Right Ear only



Show Questions for this Player: 11

Geography,

- 65. Was the topic of Speaker's lecture on **Geography**? MS
 Yes | No | Don't Know
- 66. Had the teacher begun with the **advanced** topics of the subject? MS
 Yes | No | Don't Know
- 67. Did speaker tell that the earth surface has never **changed**? MI
 Yes | No | Don't Know
- 68. Did speaker tell that Geography covers the **study of nature**? MI
 Yes | No | Don't Know
- 69. Did speaker tell that geography has **two** main branches? DTS
 Yes | No | Don't Know
- 70. Did speaker tell that Geography includes the study of the **relationship between the neighboring countries**? DTS
 Yes | No | Don't Know
- 71. Considering the content spoken by the speaker, is the talk **suitable for the beginners** to learn the subject? DTI
 Yes | No | Don't Know
- 72. According to the speaker does our way of living **impact** the planet? DTI
 Yes | No | Don't Know

Personal Experience,

- 169. Did speaker talk about his **dance** skills? MSI
 Yes | No | Don't Know
- 170. Is speaker **good** in the Spanish language? MSI
 Yes | No | Don't Know
- 171. Is the speaker **multi-lingual**? MI
 Yes | No | Don't Know
- 172. Did user work for a role that was **not** of his interest? MI
 Yes | No | Don't Know
- 173. Does the agency in South America runs **commercial** projects? DTSI
 Yes | No | Don't Know
- 174. Was it an **out of options** for the speaker to involve in the building projects? DTSI
 Yes | No | Don't Know
- 175. Did speaker worked as a **doctor**? DTI
 Yes | No | Don't Know
- 176. Does speaker has proficient **English** skills? DTI
 Yes | No | Don't Know

Player: 12, **Type:** Female voice is continuous and male voice is in chunks, **Presentation:** Female voice to the left ear only and male voice to the Right Ear only



<http://bit.ly/2DJNUr9>

Show Questions for this Player: 12

Education Museum,

- 49. Did the speaker talk about a **Library**? MS
 Yes | No | Don't Know
- 50. Did the speaker talk about **the opening and closing time** of the place MS
 Yes | No | Don't Know
- 51. Does the place **entertain students visits**? MI
 Yes | No | Don't Know
- 52. Does the place remain open for a whole week? MI
 Yes | No | Don't Know
- 53. Does the tour guide meet the visitors in **Car Park**? DTS
 Yes | No | Don't Know
- 54. Is the entrance of the place **small**? DTS
 Yes | No | Don't Know
- 55. Does the place close **early** on Tuesday than Monday DTI
 Yes | No | Don't Know
- 56. Does the place observe a **holiday on Christmas** DTI
 Yes | No | Don't Know

Travel Guide,

- 177. Did speaker talk about **life in a University**? MSI
 Yes | No | Don't Know
- 178. Did speaker give an **distance estimate** to reach another place from one place? MSI
 Yes | No | Don't Know
- 179. Does the destination has **road connectivity** with present location? MI
 Yes | No | Don't Know
- 180. Does there operate **public transport** to reach to destination? MI
 Yes | No | Don't Know
- 181. Was it the **Sydney** city for which speaker gave distance information? DTSI
 Yes | No | Don't Know
- 182. Did speaker suggested to **travel by air**? DTSI
 Yes | No | Don't Know
- 183. Is there the **taxi service available** at the location? DTI
 Yes | No | Don't Know
- 184. Was **Airport** the present location of the speaker? DTI
 Yes | No | Don't Know

Player: 13, **Type:** Both the voices are continuous, **Presentation:** First the Female voice and then the male voice with no Effect



<http://bit.ly/2TkTyog>

Show Questions for this Player: 13

University Faculty,

- 193. Did speaker talk about a **university**? MIS
 Yes | No | Don't Know
- 194. Does speaker discuss the **course content of a subject**? MIS
 Yes | No | Don't Know
- 195. Does University have only **one** faculty? MII
 Yes | No | Don't Know
- 196. Are the students **prohibited** from contacting the faculty staff? MII
 Yes | No | Don't Know
- 197. Does the speaker work as a **lecturer** in the University? DTS
 Yes | No | Don't Know
- 198. Does faculty have **five** divisions under a dean? DTS
 Yes | No | Don't Know
- 199. Does University offer education in **Technology**? DTI
 Yes | No | Don't Know
- 200. Does speaker work in the **Law** Faculty? DTI
 Yes | No | Don't Know

Hotel Facilities,

- 201. Did speaker talk about **hotels**? MIS
 Yes | No | Don't Know
- 202. Did speaker discuss the **security issues** related to place? MIS
 Yes | No | Don't Know
- 203. Do all the luxurious hotels **provide all the basic needs** of their customers? MII
 Yes | No | Don't Know
- 204. Did speaker try to **utilize the imagination capabilities** of listeners to explain the problem? MII
 Yes | No | Don't Know
- 205. Is the Travel and Public Relations a **local** research Company? DTS
 Yes | No | Don't Know
- 206. Did speaker discuss the **luxury end** of the hotel? DTS
 Yes | No | Don't Know
- 207. Was it an **evening** time when speaker gave his talk? DTI
 Yes | No | Don't Know
- 208. Was it the **first** lecture by the speaker on hospitality and tourism? DTI
 Yes | No | Don't Know

Player: 14, **Type:** Both the voices are continuous, **Presentation:** First the Female voice and then the male voice with double play-rate



Show Questions for this Player: 14

Handedness,

- 209. Did speaker talk about **disability**? MS
 Yes | No | Don't Know
- 210. Did speaker **suggest an article** to the sports people for reading? MS
 Yes | No | Don't Know
- 211. Does speaker **have an interest** in reading about Music? MI
 Yes | No | Don't Know
- 212. Is the article **suitable** to read in forming strategies in games? MI
 Yes | No | Don't Know
- 213. Is the speaker **right** handed? DTS
 Yes | No | Don't Know
- 214. Did speaker **read** an article of a sports psychologist? DTS
 Yes | No | Don't Know
- 215. Did the article of psychologist **only** cover sports? DTI
 Yes | No | Don't Know
- 216. Is the psychologist a **female**? DTI
 Yes | No | Don't Know

Buy Bicycle,

- 217. Did speaker give **tips on buying a bicycle**? MSI
 Yes | No | Don't Know
- 218. Is the single speed type of bicycle **not reliable**? MSI
 Yes | No | Don't Know
- 219. Does the speaker give tips **weekly** to buy things? MI
 Yes | No | Don't Know
- 220. Is it a **simple task** to buy a bicycle that doesn't require planning? MI
 Yes | No | Don't Know
- 221. Is the range of bicycle **limited**? DTSI
 Yes | No | Don't Know
- 222. Is the maximum price of bicycle is **\$5000**? DTSI
 Yes | No | Don't Know
- 223. Has speaker ever provided tips **other than the bicycle**? DTI
 Yes | No | Don't Know
- 224. Can one buy a bicycle in **\$45**? DTI
 Yes | No | Don't Know

- Would you prefer multiple information Communication in parallel?
 - Always
 - Sometimes, depending on information type I am interested to listen
 - No, I would prefer to listen to information with high speed as received in palyer 14
 - Never

- Can you please share us your experience in using this system (Optional)?

Submit Response

Appendix E

Ethics Application for Study IV

UTS Creativity and Cognition Studios

3-page Ethics Approval Application

From: *Muhammad Abu ul Fazal, Dr. Sam Ferguson, Dr. Andrew Johnston*

Project Number: 2018-7, HREC 2013000135

1. Title

Listening to Various Combinations of Concurrent Speech-based Information Streams

2. Aims

Studying users' experience of Listening to Various Combinations Speech-based Information Streams would determine:

- the suitability of various combinations
- performance in various combinations of streams
- effort/cognitive in various combinations of streams

3. Methodology

The following methods will be applied:

- A number of audio stimuli will be generated to reflect various combination of speech-based information streams.
- A web-based application will be developed using PHP, MySQL, HTML CSS, Bootstrap and Javascript to play the stimuli and record user's response in the database
- The experiment would be conducted in the Creativity & Cognition Studios where users will listen to the stimuli using standardized equipment (headphones and PC).

4. Significance

The analysis will identify which combination of concurrent speech-based information streams, if any, is most suitable for the users to listen and understand.

5. Number of participants and justification of numbers

30

Since our experiment would investigate the suitability of various combinations of concurrent speech-based information streams, therefore, a minimum of 30 motivated users would be required to participate in the experiment.

6. Selection/exclusion criteria

The users would mostly be from the Faculty of Engineering and IT, UTS. Since all of the information will be in English native English speakers with no significant hearing impairment will be given preference to participate in the experiment to avoid language issues contributing to poor comprehension.

7. Children under 18 years of age will participate in the evaluation.

NO

8. Procedures

Users will be briefed about the system and given instructions on how to use it. The system will be designed to be user-friendly so that users should not find any difficulty in using it.

- Users will optionally provide their profile information to the system that includes Name, Email, Age, Gender, Primary Language, Highest Qualification, Profession, Hearing Impairment, and Visual Impairment:
- Each user will randomly be assigned with 16 out of 900 stimuli
- Each user will listen to stimuli on one of the three MAC computers and three DT 770 Pro headphones.
- After hearing each stimulus, experience will be shared by the users on a simple html form
- All the responses will be stored in a database for post analysis

9. Time commitment for participants

The length of each stimulus is 2 minutes. That means a user would spend around 32 minutes to listen to all the stimuli and 18 minutes to share the experience. User may also take 10 minutes break.

Accumulatively, 60 minutes will be required from a user to participate in the experiment.

10. Location of research

The experiment will be set up in the Creativity and Cognition Studios (CCS).

11. Consent procedures

Signed consent sheet (see attached)

12. Additional Risks (additional to those noted in the CCS Generic Approval)

none

13. Strategies to cope with risks mentioned in 12.

none

14. Funding source(s) & potential conflicts of interest

No conflicts of interest.

15. Strategies to cope with any conflicts of interest identified in 14.

Not applicable.

15. Other issues

No other issues perceived as being problematic.

*Number obtained from CCS Ethics Administrator prior to completing form.

UTS: IT: CREATIVITY & COGNITION STUDIOS

AUDIENCE STUDIESUTS HREC REF 2013000135

Listening to Various Combinations of Concurrent Speech-Based Information Streams, CCS Internal Application number 2018-7

GENERAL INFORMATION

WHO IS DOING THE RESEARCH?

My name is Muhammad Abu ul Fazal and I am a student at UTS. My supervisors are Dr. Sam Ferguson, and Dr. Andrew Johnston.

WHAT IS THIS RESEARCH ABOUT?

This research is to investigate users experience against *various combinations of* concurrent voice-based audio streams played simultaneously.

IF I SAY YES, WHAT WILL IT INVOLVE?

I will ask you to

- listen 16 audio stimuli on a Mac computer using headphones
- give response on html-based experience range scale after listening to each stimuli
- optionally give your basic profile information that includes: Name, Email, Age, Gender, Primary Language, Highest Qualification, Profession, Hearing Impairment, and Visual Impairment

ARE THERE ANY RISKS?

There are minimal risks involved in this research. The volume of the audio streams played will not exceed normal conversational levels.

WHY HAVE I BEEN ASKED?

Because you

- are a native English speaker; or,
- can hear and understand English language without any difficulty; and
- do not have any significant hearing impairment,

DO I HAVE TO SAY YES?

You don't have to say yes.

WHAT WILL HAPPEN IF I SAY NO?

Nothing. I will thank you for your time so far and won't contact you about this research again.

IF I SAY YES, CAN I CHANGE MY MIND LATER?

You can change your mind at any time and you don't have to say why. I will thank you for your time so far and won't contact you about this research again.

WHAT IF I HAVE CONCERNS OR A COMPLAINT?

If you have concerns about the research that you think I or my supervisor can help you with, please feel free to contact us on Muhammad.AbuUIFazal@student.uts.edu.au, Samuel.Ferguson@uts.edu.au, andrew.johnston@uts.edu.au

If you would like to talk to someone who is not connected with the research, you may contact the Research Ethics Officer via Research.Ethics@uts.edu.au, and quote this number: 2013000135.

This study has been approved by the University of Technology, Sydney, Human Research Ethics Committee.

If you have any complaints or reservations about any aspect of your participation in this research that you cannot resolve with the researcher, you may contact the UTS Ethics Committee through the Research Ethics Officer at UTS Broadway, Building 1, Level 14; or 9514 9772; or Research.Ethics@uts.edu.au. Please quote the UTS HREC reference number.

Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

UTS: CREATIVITY & COGNITION STUDIOS

CONSENT FORM UTS HREC REF NO 2013000135

Listening to Various Combinations of Concurrent Speech -based Information Streams, CCS Internal Application number 2018-7

I _____ (participant's name) agree to participate in the research project *Listening To Various Combinations of Concurrent Speech-based Information Streams* (HREC 2013000135 project number 2018-7) being conducted by *Muhammad Abu ul Fazal, Dr. Sam Ferguson and Dr. Andrew Johnston* of the Creativity and Cognition Studios at the University of Technology, Sydney.

I understand that the purpose of this study is to share my experience of Listening to *Various Combinations of Concurrent Speech -based Audio Streams* simultaneously.

I understand that I have been asked to participate in this research because I

- am a native English speaker; or,
- can hear and understand English language without any difficulty
- do not have any significant hearing impairment,

I understand that my participation in this research will include

- listening to 16 audio stimuli on a Mac computer using headphones
- giving response on html-based experience range slider after each stimuli
- optionally, giving my basic profile information that includes: Name, Email, Age, Gender, Primary Language, Highest Qualification, Profession, Hearing Impairment, and Visual Impairment

The system will store my profile information and response to system for post analysis.

I am aware that I can contact Muhammad Abu ul Fazal, Muhammad.AbuUIFazal@student.uts.edu.au if I have any concerns about the research. I also understand that I am free to withdraw my participation from this research project at any time I wish, without consequences, and without giving a reason.

I agree that the research data gathered from this project may be published in a form that does not identify me in any way.

_____/_____/_____
Signed by

_____/_____/_____
Witnessed by

NOTE:

This study has been approved by the University of Technology, Sydney, Human Research Ethics Committee.

If you have any complaints or reservations about any aspect of your participation in this research that you cannot resolve with the researcher, you may contact the UTS Ethics Committee through the Research Ethics Officer at UTS Broadway, Building 1, Level 14; or 9514 9772; or Research.Ethics@uts.edu.au. Please quote the UTS HREC reference number.

Any complaint you make will be treated in confidence and investigated fully and you will be informed of the outcome.

Appendix F

Interface with Selected Playable Audio Files URLs, and Questionnaire for Study IV

All the web URLs mentioned in the interface are playable stimuli (open in web browser)
representing relevant designs.

Profile

You need to Login / Register first.

 Select Age 

 Select Gender 

 Select Primary Language 

 Qualification 

 Profession / Discipline 

 Select Country 

 Select Your Mood 

By clicking the check box I agree to participate in the this research experiment "Listening to Various Combinations of Concurrent Speech -based Information Streams (HREC 2013000135 project number 2018-7)". To view consent form [click here](#) (PDF file opens in another Tab).

Instructions



Instructions

Your 45-50 minutes long participation will include:

- listening to 16 audio stimuli on an iMac using headphones
- giving experiential responses on an interactive HTML form after listening to every 2 minutes long stimulus
- <https://youtu.be/oxYmUScrtm8>

I have read the Instructions.

Player: 1

1. Weight Subscale: From each pair select a dominating factor.

Temporal Demand | Frustration

Frustration | Effort

Effort | Performance

Physical Demand | Temporal Demand

Mental Demand | Effort

Effort | Physical Demand

Performance | Temporal Demand

Performance | Frustration

Temporal Demand | Effort

Mental Demand | Physical Demand

Physical Demand | Performance

Frustration | Mental Demand

Physical Demand | Frustration

Performance | Mental Demand

Temporal Demand | Mental Demand

2. Raw Rating Subscale: From each factor select the impact from low to high.

- 20. Effort How hard did you have to work to accomplish your level of performance?

Lowest -- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -- Highest

- 22. Frequent How frequently would you be using this combination of streams?

Lowest -- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -- Highest

- 21. Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Lowest -- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -- Highest

- 23. Use How much did you like this combination?

Lowest -- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -- Highest

- 18. Temporal Demand How hurried or rushed was the pace of the task?

Lowest -- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -- Highest

- 16. Mental Demand How mentally demanding was the task?

Lowest -- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -- Highest

- 19. Performance How successful were you in accomplishing what you were asked to do?

Lowest -- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 -- Highest

Sample Combination Stimuli used in Study 4

- Monolog with Interview: <http://bit.ly/2ToFcDt>
- Monolog with Commentary: <http://bit.ly/2OMtUVS>
- Monolog with News Headlines: <http://bit.ly/2OLbmFn>
- Monolog with Song: <http://bit.ly/2A5vuwC>
- Monolog with Music: <http://bit.ly/2zhS7hM>
- Interview with Commentary: <http://bit.ly/2K9drKz>
- Interview with News Headlines: <http://bit.ly/2OMIVqF>
- Interview with Song: <http://bit.ly/2TnRvQ2>
- Interview with Music: <http://bit.ly/2Q94ImQ>
- Commentary with News Headlines: <http://bit.ly/2Tlzm5G>
- Commentary with Song: <http://bit.ly/2BcJeYv>
- Commentary with Music: <http://bit.ly/2PY9jCP>
- News Headlines with Song: <http://bit.ly/2TjNtIA>
- News Headlines with Music: <http://bit.ly/2DJqyC4>
- Song with Music: <http://bit.ly/2zdRJ3U>
- Baseline: <http://bit.ly/2QPWZSa>

Appendix G

Publication 1

M. A. u. Fazal and M. Shuaib Karim, "Multiple information communication in voice-based interaction," in *Advances in Intelligent Systems and Computing*. Springer, pp. 101–111

Multiple Information Communication in Voice-based Interaction

Muhammad Abu ul Fazal (✉) and Muhammad Shuaib Karim

Department of Computer Sciences
Quaid-i-Azam University, Islamabad, Pakistan
fazalsidhu@yahoo.com, skarim@qau.edu.pk

Abstract. Ubiquitous Computing has enabled users to perform their computer activities anytime, anyplace, anywhere while performing other routine activities. Voice-based interaction often plays a significant role to make this possible. Presently, in voice-based interaction system communicates information to the user sequentially whereas users are capable of noticing, listening and comprehending multiple voices simultaneously. Therefore, providing information sequentially to the users may not be an ideal approach. There is a need to develop a design strategy in which information could be communicated to the users through multiple channels. In this paper, a design possibility has been investigated that how information could be communicated simultaneously in voice-based interaction so that users could fulfill their growing information needs and ultimately complete multiple tasks at hand efficiently.

Keywords: Voice-based Interaction, Multiple Information Broadcast, Multiple Voices, Information Design

1 Introduction

In this information age which is highly influenced by technology, people have many computing devices and associated interaction modes to fulfill information needs and perform desired tasks conveniently from anywhere [1]. For example, mobile telephony has become an essential tool [2] that humans carry with them almost all the time. It is playing a significant role to access information on the go by either interacting visually or by using voice-based interaction. A voice-based interaction is a mode where users are provided with the facility to interact with the system using 'voice'.

The motivation of using voice to interact with the system is an old concept which can be associated with Ali Baba's 'Open Sesame' and earlier science fiction movies. The voice-based interaction method enables the user to interact with the system in immersive environment [3]. *Voice User Interfaces (VUI)s are user interfaces using speech input through a speech recognizer and speech output through speech synthesis or pre-recorded audio* [4].

The voice-based interaction enables users to conveniently interact with the system in the hand busy or the eye busy environment. This mode is also an alternative for the visually impaired users to interact with systems. According to world health organization [5], it is estimated that there are 285 million people who have visual impairments.

Humans are capable of listening and comprehending multiple information simultaneously through their auditory perception, but presently, voice-based interaction design is providing sequential interaction approach which is somehow under-utilizing the natural human perception capabilities [6]. Since the voice-based interaction is sequential therefore system provides only a limited amount of information each time, which makes it hard for the users to get an overview of the information, particularly in the case of assistive technology used by the visually impaired users [8].

One of the main goals in information design is rapid dissemination with clarity. Since users have growing information needs[9] , therefore, it must be efficiently designed, produced and distributed, so that users could quickly interpret and understand it using their auditory capabilities. If we critically look contemporary implementations, then there arises a question whether present sequential information designs are utilizing the human auditory capabilities in voice-based interaction effectively & optimally or not?

Rest of the paper is organized as follows. Next section describes Literature Review. The limited exploitation of human auditory perception is discussed under the section Auditory Perception's Exploitation Gap. The concept of communicating multiple information using multiple voice streams is discussed under Motivating Scenario section. Then, based on the motivating scenario, an experiment is described in detail under Experiment Section. Conclusion and future work is discussed under Conclusion and Future Work section.

2 Literature Review

Voice-based interaction s often used in today's computing era that is ubiquitous in nature. Over the Web, efforts have also been made to realize voice-based user agents such as voice-based Web browsers under the Spoken Web Initiative [10]. That would benefit people who are unable to conveniently use the internet due to various reasons including low literacy, poverty, and disability.

There are many other uses of voice in system interaction like in e-learning system, the aural access is being provided as a complimentary method to the visual-only content [11]. Numerous interactive voice response applications are developed to provide important information to the targeted users, particularly the illiterate users. Interactive voice application 'Avaaj Otalo' [12] provides essential information to the low literate rural farmers. Using this application, farmers can ask questions, and browse stored responses on a range of agricultural topics.

From the user side, Lewis suggested that user system interaction performance is affected by the users' characteristics like physical, mental, and sensory abilities [13]. For voice, the main sensory capability is auditory acuity. The American Speech-Language-Hearing Association has identified central auditory process as the auditory system mechanisms and processes responsible for the following behaviors [14]:

- Sound localization and lateralization, i.e. users are capable of knowing the space where sound has occurred
- Auditory discrimination, i.e. user has the ability to distinct one sound from another

- Auditory pattern recognition, i.e. user is capable of judging differences and similarities in patterns of sounds
- Temporal aspects, i.e. user has abilities to sequence sounds, integrate a sequence of sounds into meaningful combinations, and perceive sounds as separate when they quickly follow one another
- Auditory performance decrements, i.e. user is capable of perceiving speech or other sounds in the presence of another signal
- Auditory performance with degraded acoustic signals, i.e. user has the ability to perceive a signal in which some of the information is missing

Humans are able to listen to the sound whose frequency varies between 16 Hz to 20KHz. In order to perceive the two frequencies separately the width of the filters, also called 'critical band', determines the minimum frequency spacing. It would be difficult to separate two sounds if it falls within the same critical band. Besides frequency, other important perceptual dimensions are pitch, loudness, timbre, temporal structure and spatial location.

Humans are capable of focusing their attention to an interested voice stream if they perceive multiple information simultaneously as reflected in experiment discussed in this paper. For attention user adopts two kinds of approaches, one is overt attention and second is covert attention. In covert attention the region of interest is in the periphery. So, if a user is listening multiple voices, he may be interested in focusing the voice provided to him in the periphery. The regions of interest could be four to five [15]. For selection and attention in competing sounds, it is an important consideration for the listener that how auditory system organizes information into perceptual 'streams' or 'objects' when multiple signals are sent to the user. In order to meet this challenge, auditory system groups acoustic elements into streams, where the elements in a stream are likely to come from the same object [Bregman, 1990].

A few research studies exist on communicating information using voice simultaneously. The experiments have been conducted particularly in the case of visually impaired persons. According to Guerreiro, multiple simultaneous sound sources can help blind users to find information of interest quicker by scanning websites with several information items [16]. Another interesting work where Hussain introduced hybrid feedback mechanism i.e. speech based and non-speech based (spearcon) feedback to the visually impaired persons while they travel towards their destination [17]. The feedback mode alters between above two modes on the basis of the frequency of using the same route by the user and representativeness of the same feedback provided to the user. The experiment conducted by the researcher reflects that hybrid feedback is more effective than the speech only feedback and non-speech only feedback. In another study for blinds to understand in a better way the relevant source's content, Guerreiro and Goncalves, established that use of two to three simultaneous voice sources provide better results [18]. The increasing number of simultaneous voices decreases the source identification and intelligibility of speech. Secondly, the author found that the location of sound source is the best mechanism to identify content.

Above mentioned behavioral characteristics and research work suggest that human auditory perception has remarkable capabilities which are somehow not fully exploited

in the contemporary implementations of the voice-based human-computer interaction, particularly for sighted users.

3 Suggested Improvements

Contrary to the voice-based interface, the visual interface provides multiple information to the user in many ways such as using overlays [19]. Figure 1 is a Facebook wall of a user where multiple information is being communicated simultaneously. One overlay is providing the facility to view the messages being received in the conversation. Another overlay at the top is showing notifications. The right side pane is showing the activities of fellows. The left side pane is displaying his favorites and other useful stuff. And as soon as the mouse is rolled over to the text Farrukh Tariq Sidhu the preview of Farrukh's wall gets displayed in another layer. If the user is interested in the additional information provided through overlay the user may go with it otherwise ignore the overlay and would stay on the main screen.

The same design technique may be adopted in voice response system to communicate multiple information simultaneously because auditory system is capable of performing filtration of received sounds and allows the user to ignore the irrelevant noise and concentrate on important information [7].

In next section, we have discussed a scenario where multiple voice streams can help users to fulfill their information needs.



Fig. 1. An example of overlay in Graphical User Interface

4 Motivating Scenario: Listening Multiple Talk Shows

Daily, in prime time i.e. 8:00 pm to 9:00 pm various news channels air talk shows focusing different topics with different participants and hosts. People working in offices in evening or night shifts usually watch these programs live using video streams provided

by news channels, if they are free to do so at the desk. If users are busy in official work or their computer screen is occupied for another task they may prefer to listen to live audio stream from relevant channels website.

Users may be interested in listening to more than one talk shows at the same time. For an example, a person is interested in listening to the talk show 'Capital Talk' at 'Geo News' and also interested in listening to 'Off the Record' played on another channel 'ARY News'. The first talk show Capital Talk is discussing the current situation arisen due to the heavy floods whereas the second program is discussing the political scenario in Pakistan. The user is mainly interested in listening to the program discussing the political situation but also wants to know the key facts or get an overview about the flood situation being discussed in Capital Talk show.

In this perspective, user's multiple information needs may be fulfilled using multiple information communication simultaneously. In this case, information seeking could be possible in a way that a user opens two web browsers and play both the audio streams simultaneously and listens both the program in parallel. This could be challenging and complex task for the user. The listening complexity may be reduced by keeping one streams volume low but audible and keep the main programs voice normal so that user could keep the focus on primary program. The high volume is expected to help him to keep the focus on the main program while the secondary low volume would continuously give him the feedback or glimpse that what is going on in the other program. Using this approach user might not miss the content of the program in which user is mainly interested and also get an overview of the secondary program.

This approach of playing multiple audio streams in parallel may be extended to more than two audio streams where information like a commentary on cricket match could also have listened.

In order to meet this challenge, we have framed following three research questions which we are trying to answer by conducting a series of experiments.

- How many voice streams can optimally be played to users for communicating information simultaneously?
- What could be the optimal auditory perceptual dimensions' settings of streams for better discrimination between voice streams?
- What scenarios / challenges users can face in multiple information communication?

5 Experiment

In this experiment, an audio bulletin was built wherein the voice-based information was designed in a way that two different voice streams (using female and male voices) were played simultaneously. The female voice stream was of BBC Urdu's renowned TV presenter 'Aaliya Nazki' and reporter 'Nasreen Askri' whereas male voice was of another BBC Urdu's TV presenter 'Shafi Taqqi Jami'.

5.1 Experiment Design and Settings

In order to build an audio bulletin, two different video bulletin of BBC Urdu's program 'Sairbeen' were selected. Sairbeen is one of the renowned news bulletins that includes

worldwide reports, expert opinions, public opinions, features on interesting topics and current affairs. This program is very popular among the public. These video bulletins were converted into two audio files of wav format. Each audio file consisted of three different news stories. From the first audio file which was in Aalia Nazki's voice, a detailed news about an exhibition scheduled to be held in Mohatta palace was selected. And from the second audio file of Shafi Taqi Jami, the main headlines of all three news were selected. These three headlines were further broken into three audio files. Each audio file played a news headline.

In order to play these news streams a different information design strategy i.e. multiple information communication simultaneously was used. In this bulletin, the detailed exhibition news was set to play continuously throughout the bulletin in a female voice. This voice stream was termed as a primary voice in the experiment. Moreover, while keeping the primary voice in playing mode the other three news stored in three audio files were also played after periodic intervals of 10 seconds. This voice considered as a secondary voice. The primary voice was set to come from left earphone whereas the secondary voice was set to come from right earphone. This approach was adopted because it was expected that playing primary and secondary voice in different earphones would bring ease for the user to discriminate both voice streams.

These two files with given information design were merged into one clip and played by writing a program in Visual Studio 2013 using C#. The total duration of this clip was 1 minute and 28 seconds. This clip was played on Dell Vostro 5560 with Core i5 processor and 4GB RAM. In order to listen to the clip, an average quality KHM MX earphone was used to listen to the clip.

The experiment was conducted on people ranging from 20 to 55 years including both males and females. Total 10 users participated in this experiment out of which 6 were male and 4 were female. The experiment was conducted at random places without considering whether the environment / surrounding was fully quiet or not.

In order to judge the behavior of users in the experiment, a questionnaire was prepared. The interviewees were first briefed about the audio playing mechanism in this experiment. They were told about both the primary and the secondary voices. Before they started to listen to the audio clip they were given an overview of the questionnaire so that they could grab the information accordingly. The questionnaire aimed to establish whether a listener could notice, focus and comprehend multiple information simultaneously or not. It also helped to gauge the notice, selection and attention behavior of the user. In order to facilitate and reduce the memory load, participants were given maximum three choices to select one from.

5.2 Results

Most of the users were able to answer the questions correctly which were asked to find out, whether they could hear both the sounds simultaneously or not. And when they were asked about the perceptual and observational question all of the participants found voice streams audible, discriminable when played together.

Following is the response of users for each question asked in the questionnaire.

i. Could you hear the primary voice presenting documentary? From all participants, 80% of the users told that the primary voice presenting documentary was clear.

The remaining 20% users who although said that they were able to listen to the primary voice but remarked that it was loud and shrilling so could further be improved.

ii. What was the topic of primary voice? All the participants rightly told the topic of primary voice i.e. Exhibition.

iii. Where was the exhibition scheduled to hold? Fifty percents of the users could not answer it correctly. Remaining those who answered correct, guessed it using their prior knowledge. The use of user's existing knowledge behavior would fully be investigated in upcoming series of experiments.

iv. What was the venue name? All the participants correctly answered the venue name of exhibition i.e. 'Mohatta Palace'.

v. Could you please tell us more about the exhibition documentary? In order to judge users' comprehension, they were asked to describe what they listened in the exhibition documentary. All the users were able to describe the documentary and gave the overview in broken words. These words were kind of keywords in the documentary that users used.

It is observed that though users lost some amount of information while focusing on secondary voice but they still grasped the documentary very well and where there was an information gap they filled it with their existing knowledge.

vi. Could you notice the secondary voice? Yes, all users were able to notice the secondary voice in the presence of primary voice.

vii. Were you able to distinguish secondary voice in the presence of primary voice and vice versa? 70% users stated that they found no difficulty to distinguish secondary sound from primary voice and vice versa. The 30% users were of the view that it could further be improved.

It is learned that this easiness in discriminating both the voice streams was mainly possible, because, both sounds were coming in different ears separately and also the voice streams were uttered by different gender voices i.e. male and female. In order to make 'discrimination' more evident, other auditory dimensions could also be explored.

vii. Were you able to distinguish secondary voice in the presence of primary voice and vice versa? The 70% users stated that they found no difficulty to distinguish secondary sound from primary voice and vice versa. Remaining 30% users were of the view that it could further be improved because they missed some information while focusing a particular voice.

It is learned that this easiness in discriminating both the voice streams was mainly possible, because, both sounds were coming in different ears separately and also the voice streams were uttered by different gender voices i.e. male and female. In order the make 'discrimination' more evident, other auditory dimensions could be explored which we would do in future experiments.

viii. What was secondary voice indicating? All the users correctly answered that secondary voice was indicating news.

ix. How many times secondary voice played in different intervals? The 30% users gave a wrong answer while 70% users rightly told that it was played three times.

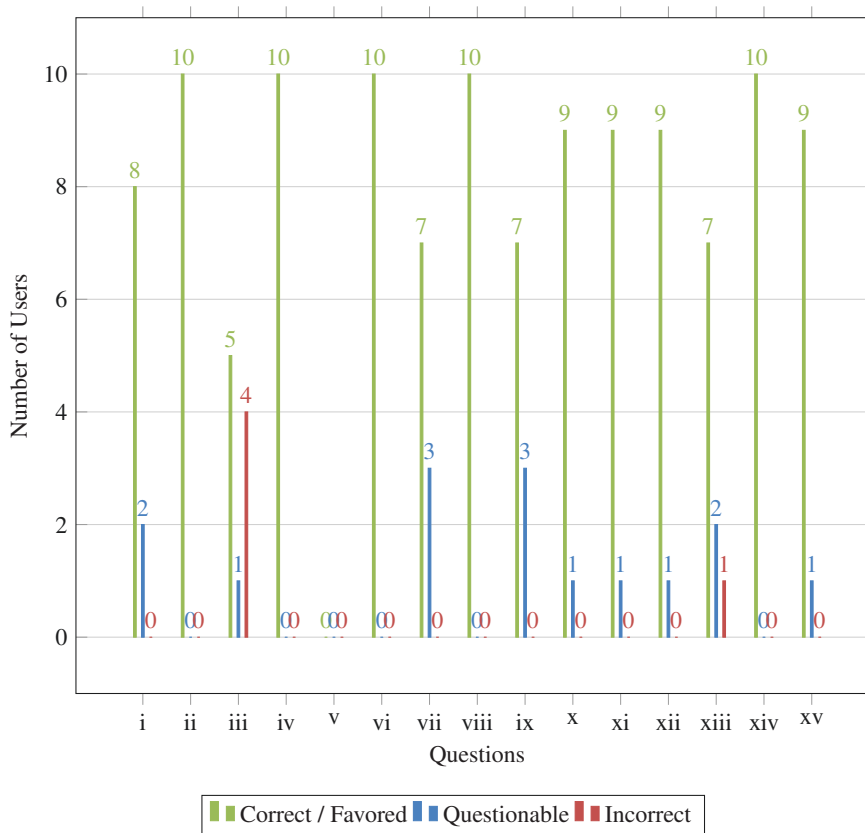


Fig. 2. Users’ response in our experiment of multiple voice-based information communication

The above bar-chart indicates the number of correct / incorrect answers by the users for each question. The question v is descriptive, therefore, not reflected in bars whereas bars in question xiii indicate the selection of interested news by the users from three headlines played to them. In question i and vii the second blue bar indicate that how many users had asked to improve the quality.

x. In the first occurrence, what was the topic of secondary voice? Among all participants, only one user couldn’t answer this question correctly.

xi. In the second occurrence, what was the topic of secondary voice? The 90% of participants correctly told the topic of the second occurrence i.e. cyber attack.

xii. In the third occurrence, what was the topic of secondary voice? Same results were witnessed as seen in above two questions.

xiii. Which was the most interesting news for you? The 70% users opted 'Data theft in Cyber Attack' remaining twenty percent of the users opted for 'Black Money in Budget' whereas only one female user showed interest in Exhibition Documentary.

xiv. Did you want to promptly listen to the detail news from any of the spoken news? As a follow-up to Question 13 when users asked to tell their intent that whether they wanted to promptly listen to the interesting news by skipping the present primary voice then 100% of the users answered 'yes'.

This is an interesting finding which provides the opportunity of applying GUI based overlay, lightbox techniques in voice-based interaction which is discussed in the previous section using the Facebook wall of a user.

xv. Did you find multiple sounds helpful in reaching multiple information quickly and Would you prefer this approach over the sequential flow of information? The 90% of the users found this quick design of delivering information helpful and said they would prefer this multiple information communication simultaneously over the sequential flow of information. From these 90% users, a few had reservations. They said, in this technique they are afraid that they might lose some important information which they would prefer to listen without any noise and disturbance. So, it could also be an interesting finding that in which contexts the multiple information communication design strategy could be applied and where it can't.

The 10% of the users who didn't give preference said they are uni-task oriented so can't prefer this approach over the sequential flow of information.

6 Conclusion and Future Work

The results of this experiment are encouraging to further explore this design approach. The results validate that multiple information communication is possible using voice in Human-machine interaction. Users showed interest in multiple information communication. They were able to discriminate the voice. Using their focus and attention abilities they were able to get multiple information meaningfully in lesser time.

We find it suitable to further investigate this information design approach. We are presently in the process to develop a software that would be able to play multiple live programs simultaneously. Each program would have its own set of controls mapped with auditory perceptions. Users would be able to set the controls, i.e. they would be able to pan the stream, make the volume low and high, change the pitch, change the rate of voice streams and much more which may help them to listen to multiple voice streams simultaneously using their focus and attention abilities. This web-based software would be used to observe the interaction behaviour of users. For example, what values they set to the control to listen to the multiple sounds?

References

1. Guo-ping Li and Guo-yong Huang. The "core-periphery" pattern of the globalization of electronic commerce. In *Proceedings of the 7th International Conference on Electronic Commerce, ICEC '05*, pages 66–69, New York, NY, USA, 2005. ACM.
2. Raman Kazhamiakin, Piergiorgio Bertoli, Massimo Paolucci, Marco Pistore, and Matthias Wagner. Having services "yourway!": towards user-centric composition of mobile services. In *Future Internet–FIS 2008*, pages 94–106. Springer, 2009.

3. Philip Kortum. *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.
4. Dirk Schnelle-Walka. I tell you something. In *Proceedings of the 16th European Conference on Pattern Languages of Programs*, page 10. ACM, 2012.
5. World Health Organization. Visual impairment and blindness. <http://www.who.int/mediacentre/factsheets/fs282/en/>, 2014. [Online; accessed 04-Jan-2016].
6. Ádám Csapó and György Wersényi. Overview of auditory representations in human-machine interfaces. *ACM Computing Surveys (CSUR)*, 46(2):19, 2013.
7. Alan Dix, Janet E Finlay, Gregory D Abowd, and Russell Beale. *Human-computer interaction*. 2003.
8. Daisuke Sato, Shaojian Zhu, Masatomo Kobayashi, Hironobu Takagi, and Chieko Asakawa. Sasayaki: Augmented voice web browsing experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pages 2769–2778, New York, NY, USA, 2011. ACM.
9. Karen Church, Mauro Cherubini, and Nuria Oliver. A large-scale study of daily information needs captured in situ. *ACM Trans. Comput.-Hum. Interact.*, 21(2):10:1–10:46, February 2014.
10. Sheetal K. Agarwal, Anupam Jain, Arun Kumar, Amit A. Nanavati, and Nitendra Rajput. The spoken web: A web for the underprivileged. *SIGWEB Newsl.*, (Summer):1:1–1:9, June 2010.
11. MPuerto Paule-Ruiz, Víctor Álvarez García, J. R. Pérez-Pérez, and M. Riestra-González. Voice interactive learning: A framework and evaluation. In *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education, ITICSE '13*, pages 34–39, New York, NY, USA, 2013. ACM.
12. Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S. Parikh. Avaaj otalo: A field study of an interactive voice forum for small agriculturers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 733–742, New York, NY, USA, 2010. ACM.
13. James R Lewis. *Practical speech user interface design*. CRC Press, Inc., 2010.
14. Ronald L Schow, J Anthony Seikel, Gail D Chermak, and Matthew Berent. Central auditory processes and test measuresasha 1996 revisited. *American Journal of Audiology*, 9(2):63–68, 2000.
15. Roxanne L Canosa. Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception (TAP)*, 6(2):11, 2009.
16. João Guerreiro. Using simultaneous audio sources to speed-up blind people's web scanning. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, page 8. ACM, 2013.
17. Ibrar Hussain, Ling Chen, Hamid Turab Mirza, Gencai Chen, and Saeed-Ul Hassan. Right mix of speech and non-speech: hybrid auditory feedback in mobility assistance of the visually impaired. *Universal Access in the Information Society*, pages 1–10, 2014.
18. João Guerreiro and Daniel Gonçalves. Text-to-speeches: evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 169–176. ACM, 2014.
19. Bill Scott and Theresa Neil. *Designing web interfaces: Principles and patterns for rich interactions*. " O'Reilly Media, Inc.", 2009.

Appendix H

Publication 2

M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, "Concurrent Voice-Based Multiple Information Communication: A Study Report of Profile-Based Users' Interaction," in *145th Convention of the Audio Engineering Society*. Audio Engineering Society, 2018



Audio Engineering Society Convention e-Brief 17

Presented at the 145th Convention
2018 October 17 – 20, New York, NY, USA

This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Audio Engineering Society.

Concurrent Voice-based Multiple Information Communication: A study report of profile-based users' interaction

Muhammad Abu ul Fazal¹, Sam Ferguson¹, Muhammad Shuaib Karim², and Andrew Johnston¹

¹*School of Software, University of Technology Sydney, Australia*

²*Quaid-i-Azam University, Islamabad, Pakistan*

Correspondence should be addressed to Muhammad Abu ul Fazal (Muhammad.AbuUlFazal@uts.edu.au)

ABSTRACT

This paper reports a study conducted with 10 blind and 8 sighted participants using a prototype system for communicating multiple information streams concurrently, using two methods of presentation. The prototype system in method one played two continuous voice-based articles diotically, differing by voice gender and content. In the second method, prototype communicated one continuous article in the female voice and three headlines as interval-based short interruptions in a male voice dichotically. In this investigation the continuous method remained more effective in communicating multiple information compared to the interval-based interruption method, and also the users who possessed at least tertiary qualification performed better in comprehending the multiple concurrent information than the non-tertiary qualified users.

1 Introduction

Humans are capable of listening and comprehending multiple information concurrently through auditory perception [1], but presently voice-based interaction designs provide a communication approach where a system delivers information sequentially to the user, underutilizing the maximum capacity of human perception capabilities [2, 3]. Since users have growing information needs, therefore, the information communication methods that transmit information to a user more efficiently should be explored fully. To achieve such efficiency multiple voice-based information communication systems can be a possibility that warrants investigation.

In concurrent multiple information streams, most of the

researchers have investigated users abilities in segregating a 'target' from the 'masker' in competing voices [4, 5, 6, 7] and established that the intelligibility of the target increases with the increase of spatial difference between the target and the maskers [8, 7]. Few researchers have looked at multiple targets, where each target masks other targets. According to Guerreiro et al., multiple concurrent sound sources can benefit blind and sighted users to find information of interest quicker by scanning several information items [9]. They ascertained that the use of two to three concurrent voice streams provides efficient results in concurrent communication [10].

This paper studies users' abilities to listen and 'comprehend both the audio streams (targets)' played concurrently. The paper is arranged as follows. The integra-

Table 1: Profile based User Groups: Created two groups based on user qualification.

Group #	Academics	Interest	No.
G1	>= Tertiary	Any	10
G2	<Tertiary	Any	8

tive analysis of the results follows a methodology of the investigation. After that, the discussion is mentioned.

2 Investigation

We investigated two approaches to simultaneous presentation of speech, 1. Continuous simultaneous presentation and, 2. Simultaneous presentation with interval-based interrupts, and compared the effectiveness of each method regarding users' comprehension of the presented information in this study. Our investigation further looked into whether the user profile, based on academic qualification, played a role in comprehending multiple voice-based information concurrently or not?

2.1 Participants

Ten blind and eight sighted users with the median age of 28 years participated in this study. All the participants were well-versed in the Urdu language, the language that was used in the prototype for content delivery. For comparison of users' performance based on their profiles, two user groups were arranged by users' qualifications that are specified in Table 1. In G1, the users possessing at least tertiary educational background were grouped, whereas those who were not holding tertiary education were grouped in G2.

2.2 Method 1 - Continuous: Stimuli & Questionnaire

In this method, the prototype system played two different TV-shows' audio content to the users concurrently. Both the audio streams were discriminable by the gender of the voice, and were set to play continuously and diotically to the users for one minute with similar sound pressure levels.

2.3 Method 2 - Continuous and Interval-Based: Stimuli & Questionnaire:

In this concurrent method following dichotic listening paradigm, a documentary stream was set to play continuously in the left ear, while the news headlines were played in the right ear after silence intervals of 20 seconds between each headline.

Measurement Users were asked to listen to the concurrent streams in both the approaches using earphones, and then answer content-based questions, with answers that would require comprehension from the audio streams. The questionnaire was prepared to assess the comprehension of content by a user from basic to advanced level in each type of concurrent method. The questionnaire also included a question that assessed the user interest in multiple information communication concurrently. A couple of the questions mentioned in the questionnaire were: • What was the topic of the male documentary? (Basic) • How many were the employees whose data was stolen? (Advance)

2.4 Protocol

An Apple MacBook Pro with left and right built-in audio speakers was used to play the prototype. Besides the built-in speakers, users were also provided with the iPhone-6 earphones to listen to the audio streams. The users used earphones to listen to the streams. Users were orally briefed about information presentation. The questions were asked in an interview form.

3 Result & Analysis

The results & Analysis are covered qualitatively and quantitatively in the following sub-sections.

3.1 Qualitative Analysis

This subsection briefly discusses the responses of the users qualitatively to share their comprehension, experiences, and expectations.

Regarding content comprehension, in the questionnaire, a descriptive (open) question, "What have you heard in the stream(s)? Please describe." was asked to assess users' comprehension of the contents subjectively. The users' response to this question is illustrated in a histogram, Figure 1, which is developed based on

the authors' subjective opinion about users' comprehension of the content. According to the histogram, 6 participants comprehended the streams content 'poor' to 'not good' level, 9 were 'moderate' and 4 users comprehended content 'good' to 'excellent' level.

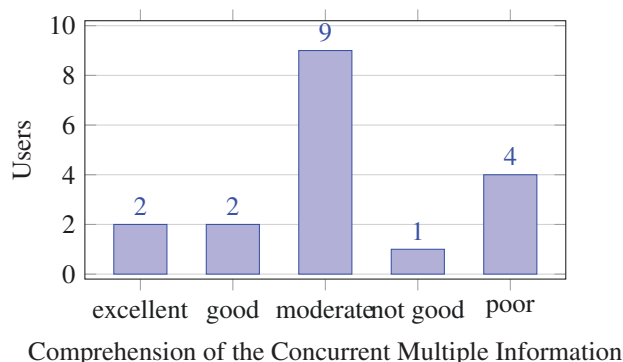


Fig. 1: Histogram, reflecting the number of users and their comprehension based on our subjective opinion and users' individual score in studies.

Moreover, regarding users' experience, reactions, and expectations, the following sub-sections concisely discuss the points reported by users based on their experience with such communication techniques. These points provide several hints and apparently open avenues for researchers to explore the directions of communicating multiple information concurrently.

Continuous vs. Interval Voices Regarding the forms of information delivery, many users reported that they were more comfortable in listening to continuous voices comparing to the interval-based communication. They suggested that continuous voices with dichotic presentation could have been more helpful. Regarding interval-based communication, they reported that the loudness of secondary voice in chunks broke the focus. However, one user found interval-based communication more helpful. He told, in dichotic listening one needs to focus on the continuous voice, whereas the mind would effortlessly catch the intervals-based voice.

Language of the Content Highlighting the significance of the language of the content, a user reported that the content in his mother tongue (the regional language) could have helped him to perform better in these studies.

Dichotic Presentation As reported by almost all the users, dichotic presentation appeared as an important factor in differentiating the streams' content from each other. Only one user argued against the dichotic presentation, and told that dichotic presentation created a focus shift issue. He argued, human mind is used to of listening to the voices in both ears (diotic), but in dichotic presentation, the voices were coming in separate ears. Therefore, the brain started to capture information randomly sometimes from the right ear and sometimes from the left ear. He concluded that both the voices should come to both ears because it's more natural (cocktail party effect) whereas dichotic presentation feels to be unnatural.

Play Controls The provisioning of the audio player controls appeared as an essential demand by the users to listen to multiple streams of information concurrently. They argued to give the audio-controls to the users so that they could set them according to their need. For example, they should be able to bring one stream's volume low, and other's high or vice-versa and also adjust the play rate of the streams according to their need and context.

Interest in the Content Some users reported that their interest in the transmitted content was an important factor in comprehending information. A user reported that he could have focused more if the audio recordings were related to the music or songs. Similarly, a user told that his interest in News helped her to score better.

Training & Practice Some of the users were of the view that practicing such system can improve comprehension in such systems. A user reported that in the start it was an unexpected behavior of information presentation for the mind but later mind started to segregate the voices easily.

Preference Lastly, answering a pertinent question, 'whether you would prefer multiple information communication concurrently over the sequential flow of information?', the equal number of users said 'Yes,' 'No,' and 'Maybe'. Many of the users who opted 'No' argued that in the concurrent form of delivery, they might miss a significant amount of information that could be a big problem when the information requirement is crucial and requires listening to it carefully and uninterruptedly. Therefore, users asked to provide

them with authority regarding player controls to decide themselves whether they want two voices streams to be played concurrently or sequentially. Those who opted 'may be' also argued that it would depend on the information-seeking context.

3.2 Quantitative Analysis

Besides the qualitative response of the users, the quantitative analysis was also performed regarding users' correct answers to the questions. A number of comparisons from different perspectives are drawn with respect to the user groups mentioned in Table 1. The first perspective is to compare the performance of the G1 with the G2, the second is to compare the performance of the blind and the sighted Users. After that, the users' performance in answering the basic and the advanced questions is compared. Finally, a comparison between both the communication method is shown.

Regarding users' group-wise comparison, Figure 2 indicates the percentages of correct and incorrect answers with respect to each group. In this study, the G1 performed better than the G2 that means the users who possessed tertiary qualification or above performed better than the users who were not holding this level of qualification. The same Figure 2 also indicates that the performance in comprehending information by the visually challenged users (VCUs) and the sighted users (SUs) remained almost the same.

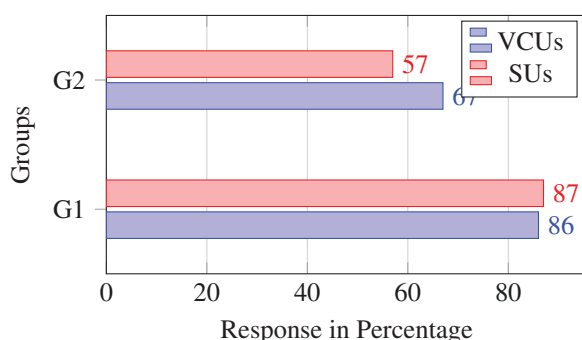


Fig. 2: Group-wise users' response in both studies. *The bars reflect the correct percentage of answers to the questions asked in both the studies.*

Regarding comprehending basic to advanced level information, the results are reflected in Figure 3. As shown, the users were able to answer more questions

correctly that were set from the basic information comparing to the questions that were set from the detailed information.

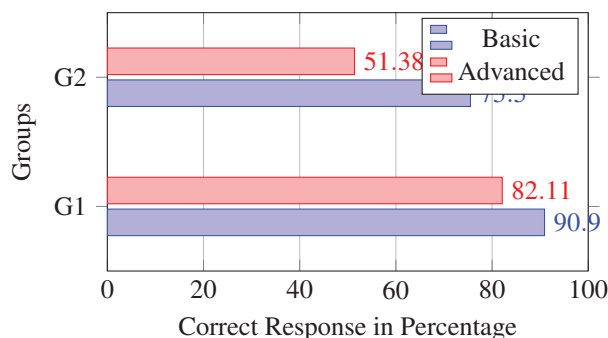


Fig. 3: Group-wise users' response in basic and advanced questions.

Regarding comparing the effectiveness of both the methods, i.e. continuous or interval-based, the continuous content delivery was comprehended better than the interval-based communication. Figure 4 indicates the results. In the continuous method, 85% questions were answered correctly, whereas in interval-based method 77% in basic types of questions and 67% in advanced types of questions were answered correctly by the users.

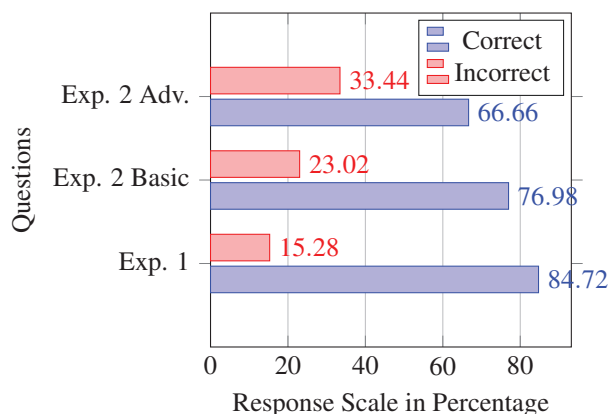


Fig. 4: Response of all users for both studies.

4 Discussion

Although some of the users manifested a disinclination towards multiple concurrent voice information communication, their scores in the questionnaire revealed

that they performed well in comprehending the information from voice streams played concurrently. Many of the participants were interested in the development of new technologies that may assist them in meeting daily challenges of extensive information.

The dichotic listening based on audio panning, as reported by most of the users, helped them in segregating both the streams from each other. Almost, all the users agreed that 'panning' was helpful in separating the content from each other. Therefore, another approach for investigation could be to deliver continuous streams dichotically.

The objective of the second approach was to imitate the GUI overlay technique. The short audio clips of useful information were passed to the user while listening to a documentary by using almost half of the audio display bandwidth. The prototype played the headlines in the right ear that otherwise remained silent throughout the clip. It was expected that this approach would be more efficient and acceptable to the users. But most of the participants found interval-based communication as a hindrance in comprehending multiple voice streams, possibly because of the non-optimal utilization of auditory bandwidth. Therefore, the identification of optimal auditory information bandwidth could be the subject of a future investigation.

We are advancing this investigation and working on carrying out a series of standardized experiments to come up with optimal audio player settings for concurrent voice streams that may help users in comprehending multiple information concurrently without having significant challenges. These experiments would help in developing a framework for communicating multiple information concurrently.

References

- [1] ul Fazal, M. A. and Karim, M. S., "Multiple Information Communication in Voice-Based Interaction," in *Multimedia and Network Information Systems*, pp. 101–111, Springer, 2017.
- [2] Csapó, Á. and Wersényi, G., "Overview of auditory representations in human-machine interfaces," *ACM Computing Surveys (CSUR)*, 46(2), p. 19, 2013.
- [3] Dix, A., Finlay, J. E., Abowd, G. D., and Beale, R., "Human-Computer Interaction," 2003.
- [4] Carlile, S. and Schonstein, D., "Frequency bandwidth and multi-talker environments," in *Audio Engineering Society Convention 120*, Audio Engineering Society, 2006.
- [5] Deroche, M. L., Culling, J. F., Chatterjee, M., and Limb, C. J., "Roles of the target and masker fundamental frequencies in voice segregation," *The Journal of the Acoustical Society of America*, 136(3), pp. 1225–1236, 2014.
- [6] Deroche, M. L. and Culling, J. F., "Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation," *The Journal of the Acoustical Society of America*, 130(5), pp. 2855–2865, 2011.
- [7] Sikström, E. and Berg, J., "Designing auditory display menu interfaces-cues for users current location in extensive menus," in *Audio Engineering Society Convention 126*, Audio Engineering Society, 2009.
- [8] Ihlefeld, A. and Shinn-Cunningham, B., "Spatial release from energetic and informational masking in a selective speech identification task," *The Journal of the Acoustical Society of America*, 123(6), pp. 4369–4379, 2008.
- [9] Guerreiro, J., "Using simultaneous audio sources to speed-up blind people's web scanning," in *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, p. 8, ACM, 2013.
- [10] Guerreiro, J. and Gonçalves, D., "Text-to-speeches: evaluating the perception of concurrent speech by blind people," in *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pp. 169–176, ACM, 2014.

Appendix I

Publication 3

M. A. u. Fazal, S. Ferguson, M. S. Karim, and A. Johnston, "Vinfomize: A framework for multiple voice-based information communication," in *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*. ACM, 2019, pp. 143–147

Vinfomize: A Framework for Multiple Voice-based Information Communication

Muhammad Abu ul Fazal
School of Software, Faculty of Engineering & IT,
University of Technology, Sydney
Sydney, NSW, Australia
Muhammad.AbuUlFazal@uts.edu.au

Shuaib Karim
Department of Computer Science, Quaid-i-Azam
University
Islamabad, Pakistan
skarim@qau.edu.pk

Sam Ferguson
School of Software, Faculty of Engineering & IT,
University of Technology, Sydney
NSW, Australia
Samuel.Ferguson@uts.edu.au

Andrew Johnston
School of Software, Faculty of Engineering & IT,
University of Technology, Sydney
NSW, Australia
Andrew.Johnston@uts.edu.au

ABSTRACT

In this paper, two studies, conducted with 10 blind and 8 sighted users, for investigating the possibilities of communicating multiple information concurrently are reported. In the first study, we concurrently played two voice-based articles in continuous form in both the ears, and in the second study, we concurrently communicated one article continuously in one ear and three news headlines as an interval-based short interruption in another ear. In the results, we first reported the participants' experience qualitatively and also shared their expectations with multiple information communication, and then based on the feedback received from the users, we proposed a framework that may help in developing systems to communicate multiple voice-based information to the users. It is expected that such information systems thus developed would provide a better user experience regarding optimal information communication to the users vis a vis their contextual and perceptual needs and limitations.

CCS CONCEPTS

• **Information systems** → **Multimedia streaming**; • **Human-centered computing** → *Empirical studies in HCI*;

KEYWORDS

Voice-based information systems; concurrent communication; framework; multiple audio players

ACM Reference Format:

Muhammad Abu ul Fazal, Sam Ferguson, Shuaib Karim, and Andrew Johnston. 2019. Vinfomize: A Framework for Multiple Voice-based Information

Communication . In *ICISDM 2019: 3rd International Conference on Information System and Data Mining (ICISDM'19), April 06–08, 2019, University of Houston, Texas, USA*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3206098.3206116>

Due to the exponential surge in information generation in the last decade, users including visually challenged persons have found it challenging to cope with this massive information delivery [15]. According to [8], humans are capable of listening and comprehending multiple information concurrently through auditory perception, but presently voice-based interaction design provides communication approach where a system delivers information sequentially to the user which underutilizes natural human perception capabilities. Since voice-based interaction is sequential, therefore, the system renders a limited amount of information in a given time, which makes it harder for the users to get more information in less time. There are some areas where the content is not highly important, user is generally interested to get gist of information from different sources, e.g. watching a documentary and getting gist from cricket commentary etc. simultaneously. In such areas support from multiple voice-based systems can be helpful.

One of the primary goals in information communication is rapid dissemination with accuracy [5]. Since users have growing information needs, therefore, it must be efficiently distributed so that users could quickly interpret and understand it [9, 17]. Interfaces to comprehend or scan information broadcast by a user in the shorter time have not been widely investigated [8], despite, users are capable of switching attention to an interested voice stream when they receive multiple information streams concurrently. This ability has been identified in renowned 'cocktail party effect' experiment [4, 16].

For selection and attention in competing sounds, it is an important consideration for the listener how the auditory system organizes information into perceptual 'streams' or 'objects' when multiple signals reach the user. The auditory system groups acoustic elements into streams where the elements in a stream come from the same object/source [1] which helps a user to segregate one stream from the others. Moreover, regarding attention, a user may adopt two kinds of approaches; one is overt attention and the other is covert attention. In covert attention, the region of interest

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICISDM'19, April 06–08, 2019, University of Houston, Texas, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6635-9...\$15.00

<https://doi.org/10.1145/3206098.3206116>

remains in the periphery, meaning that, a user can be interested in focusing on a voice heard in the periphery [2]. These behavioral characteristics imply that human auditory perception has remarkable capabilities which are somehow not fully employed in the current implementations at large scale.

Few researchers has investigated techniques to utilize this human ability, and most of the research has been carried out to aid users in segregating a 'target' from the 'masker' [3, 6, 7, 11]. The intelligibility of the target increases as the spatial separation with the maskers increases [14]. The source identification through spatial difference facilitates users in grouping and discriminating one voice from the other [?]. Additionally, few researchers have looked at multiple targets, where each target masks other targets. According to [12], multiple concurrent sound sources can benefit blind and sighted users to find information of interest quicker by scanning several information items. They also compared variations like faster speech rate against the concurrent speech and concluded that concurrent voices with speech rates slightly faster than the default rate enable significantly faster scanning for 'finding relevant content' [13].

Given the large information generation and dissemination, there is a need to communicate more than one targets concurrently. This paper has studied users' abilities to listen and 'comprehend both the audio streams (targets)' played concurrently.

1 INVESTIGATION

We investigated two different approaches, 1. Continuous... 2. Interval-based Interrupts..., for presenting multiple information concurrently, and compared the effectiveness of each method regarding users' comprehension of the presented information and experience.

1.1 Participants

Ten blind/visually challenged users (VCUs) and eight sighted users (SUs) with a median age of 28 years participated in our studies. All the participants were well-versed in the language that we used in the experiment. For the participation of VCUs, a training center responsible for educating and training the special pupils including the visually challenged persons was officially contacted and briefed about the goals of the investigation. The Institute after obtaining the consent from the visually challenged persons agreed for the participation of their students and staff.

1.2 Study 1 - Continuous: Stimuli & Questionnaire

For this study, we concurrently played audios of the two TV-shows streams in continuous and diotic form to the users using the prototype. One stream was in female voice, and the other stream was in male voice. Both the streams were obtained from a renowned media group [10]. The topic of the female voice stream was women empowerment, and for the male voice stream, it was China-Pakistan Economic Corridor (CPEC) Development. Both the articles were in the language that was generally understandable by the participants. The streams were played concurrently for one minute to the users who were asked to listen to both streams concurrently. The stimuli design for study 1 is illustrated in Figure 1 for further elaboration.

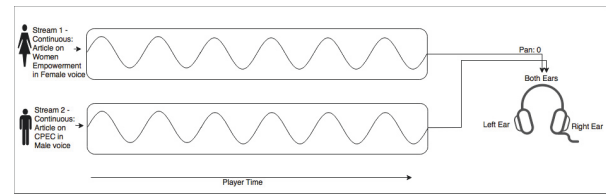


Figure 1: Stimuli design for Study 1 - Continuous

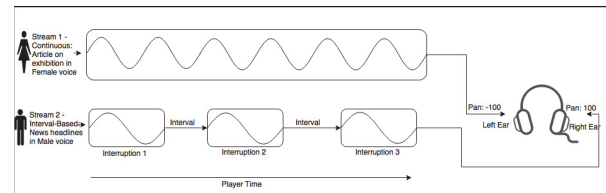


Figure 2: Stimuli design for Study 2 - Continuous and Interval-Based

Questionnaire I. In Table 1 the experiential and the basic content-based questions asked in study 1 are detailed along with the number of correct and incorrect answers by the participants graphically.

1.3 Study 2 - Continuous and Interval-Based: Stimuli & Questionnaire:

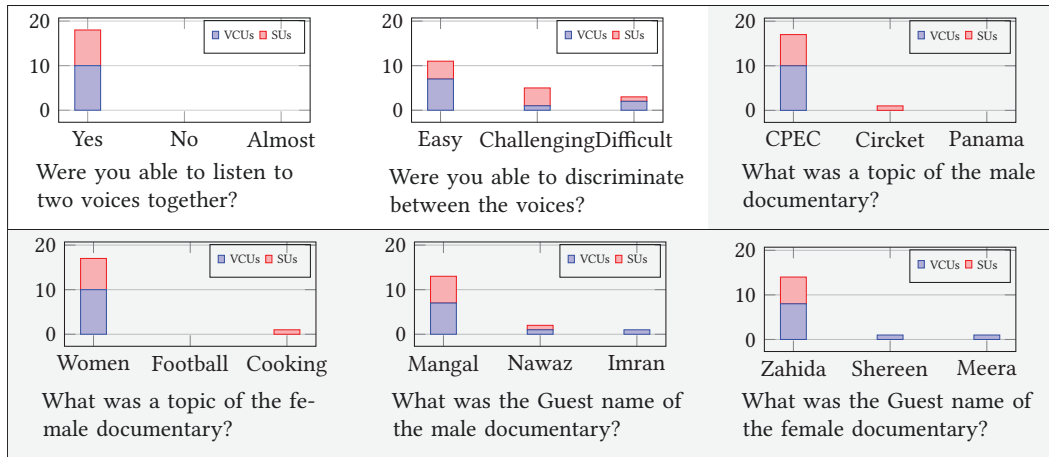
In this study, an exhibition documentary and three headlines were chosen. The exhibition documentary as stream 1 was played in the female voice, whereas the news headlines as stream 2 were played in a male voice. Based on the dichotic listening and short interruption approach, the stream 1 was played continuously in the left earphone whereas the stream 2 was played in the right earphone after the silent interval of 20 seconds between each news headline. The length of this stimulus was 70 seconds. The stimulus design for study 2 is illustrated in Figure 2 for further clarity.

Questionnaire II. The questionnaire prepared for this study is shown in Table 2, which was designed from the stated information in streams to assess the comprehension of content by a user from basic to advanced level. In basic, simple questions based on the main/prominent information were asked e.g. what was the topic of the documentary, what was the news headline indicating. In the advanced section, nine questions were asked based on the less prominent information. The questionnaire also included some questions that helped in measuring the user interest in multiple information communication concurrently.

1.4 Protocol

In these studies, no access to any of the audio player controls, e.g., volume, playback-rate, forward, and back, was provided to the users. The content was played at normal audible volume and regular playback-rate in a quiet room. An Apple MacBook Pro with left and right built-in audio speakers was used to play the prototype. Besides the built-in speakers, users were also provided with iPhone-6 earphones to listen to the audio streams. The users were first asked whether they are comfortable in using earphones

Table 1: Questionnaire 1: Questions and responses by users for Study 1. The questions in white background are experiential whereas the gray backgrounded are basic content-based questions. The first option in each content-based question is the correct answer.



or not, particularly the VCUs. The users used earphones to listen to the streams. Users were told to focus on both the voice streams. Before starting the participation, users were orally briefed about information presentation/stimuli designs. An idea about the types of questions and how to answer them was also given to users. The questions were arranged to measure the content comprehension and users experience, and were asked to the users in an interview form that helped to get a detailed response on open questions set in the questionnaire.

2 RESULT & ANALYSIS

Users response against each question in both the studies are graphically mentioned in Tables 1 and ?? using bar charts. The qualitative analyses on results are described in subsequent subsections.

2.1 Qualitative Analysis

This subsection qualitatively discusses users responses to share their comprehension and experience. Besides the content-based questions mentioned in the questionnaire, a descriptive (open) question, "What have you heard in the stream(s)? Please describe." was also asked for judging the descriptive comprehension of the contents by the users. According to authors' subjective opinion about users' comprehension, 6 participants comprehended the content of the streams from 'poor' to 'not good' scale, 9 were 'moderate' and 4 users comprehended content from 'good' to 'excellent' range.

Moreover, regarding users' experience, reactions, and expectations, following points concisely discuss the factors reported by users based on their experience with such communication techniques. These factors provide several hints and open avenues for researchers to explore the directions of communicating multiple information concurrently.

Continuous vs. Interval Voices. While sharing opinions on the forms of deliveries, some of the users reported that they were more comfortable in listening to continuous approach compared

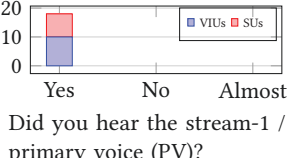
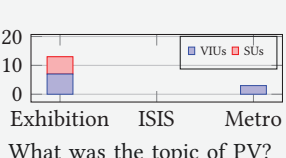
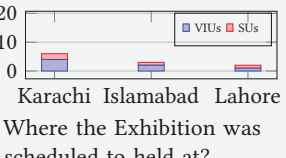
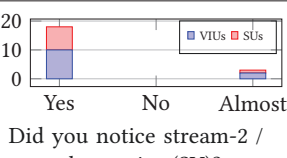
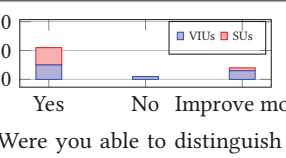
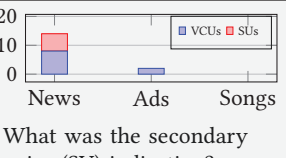
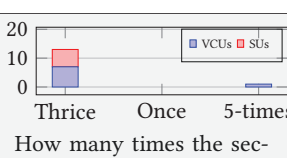
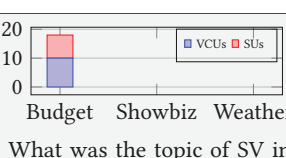
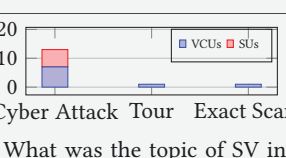
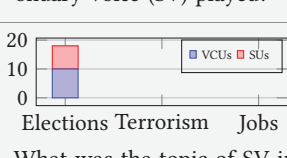
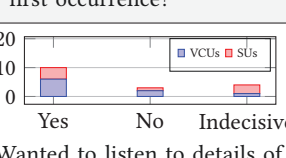
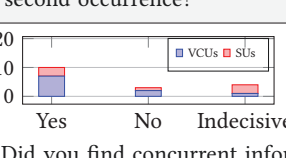
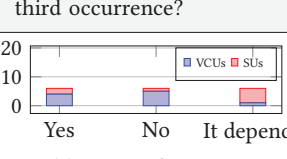
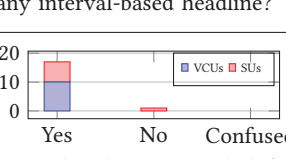
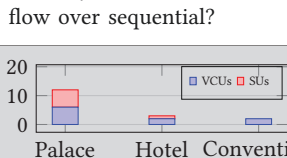
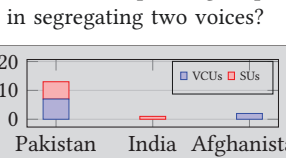
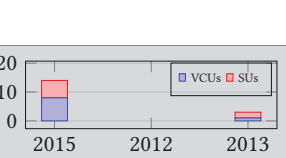
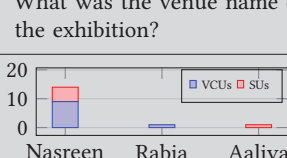
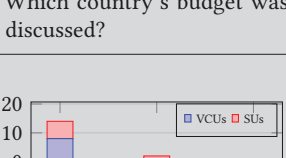
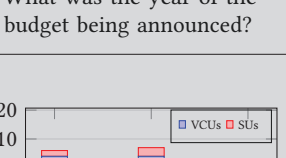
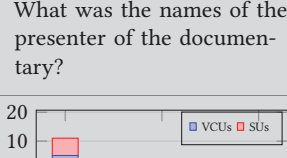
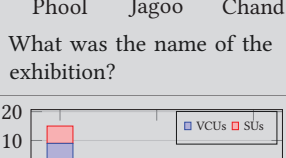
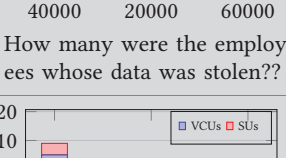
to interval-based communication. They suggested that continuous streams with dichotic presentation could have been more helpful. Regarding interval-based communication, users reported that the high sound volume of secondary voice in chunks broke their focus. However, differing to the above view, a user found interval-based communication more helpful. He told, with the help of dichotic listening and interval-based approach, one simply needs to focus on the continuous voice stream to comprehend it, and the interval-based voice stream would effortlessly be apprehended by the mind.

Moreover, as stated by many of the users that the continuous content delivery was more appropriate than the interval-based communication, the quantitative analysis validated the same. The percentage of correct answers in study one remained greater than the percentage in the second study. Also, the performance of VCUs and SUs remained similar in comprehending information in both types of approaches.

Dichotic Presentation. Dichotic presentation appeared as an important factor to segregate the streams' content from each other, as reported by almost all the users. Dichotic presentation, achieved by panning feature, helped them to localize the source of competing streams that ultimately helped them in segregating the streams. Only one user argued against the dichotic presentation and maintained that dichotic presentation created a focus shift issue. He elaborated that, in the daily routine interactions, human minds are used to of listening to voices in both ears (diotic), but in dichotic presentation, the voices were coming in separate ears. This was apparently strange behavior for the brain, and therefore, it randomly started to capture information from both streams, sometimes from right ear and sometimes from the left ear. He concluded that both the voices should come to both ears because it's more natural whereas dichotic presentation felt unnatural.

Play Controls. The provisioning of the audio player controls appeared as an essential demand by the users to listen to multiple streams of information concurrently. Users argued that for listening

Table 2: Questionnaire 2: Questions (User Experience & Basic Content-based) and responses by users for Study 2. The questions with white background are experiential whereas the gray backgrounded questions are basic content-based questions. All the dark gray are the advanced content-based questions. The first option in each content-based question is the correct answer.

 <p>Did you hear the stream-1 / primary voice (PV)?</p>	 <p>What was the topic of PV?</p>	 <p>Where the Exhibition was scheduled to held at?</p>
 <p>Did you notice stream-2 / secondary voice (SV)?</p>	 <p>Were you able to distinguish between streams?</p>	 <p>What was the secondary voice (SV) indicating?</p>
 <p>How many times the secondary voice (SV) played?</p>	 <p>What was the topic of SV in first occurrence?</p>	 <p>What was the topic of SV in second occurrence?</p>
 <p>What was the topic of SV in third occurrence?</p>	 <p>Wanted to listen to details of any interval-based headline?</p>	 <p>Did you find concurrent information comm. helpful?</p>
 <p>Would you Prefer concurrent flow over sequential?</p>	 <p>Found audio panning helpful in segregating two voices?</p>	
 <p>What was the venue name of the exhibition?</p>	 <p>Which country's budget was discussed?</p>	 <p>What was the year of the budget being announced?</p>
 <p>What was the names of the presenter of the documentary?</p>	 <p>What was the name of the exhibition?</p>	 <p>How many were the employees whose data was stolen??</p>
 <p>Which country stole the data from the internet?</p>	 <p>In which country the election was to held in?</p>	 <p>On which day the elections were held?</p>

to multiple information streams, they should be provided with audio control so that they could set the controls according to their need. For example, users should be able to bring one stream's volume low, and other's high or vice-versa and also adjust the panning and playback-rate of the streams, etc. based on their information needs and context.

Interest in the Content. Some users reported that their interest in the transmitted content was an important factor in comprehending information. A user reported that he could have focused more if the audio recordings were related to the music or songs. Similarly, a user told that his interest in News helped her to score better.

Keywords. The results demonstrated that the keywords in the content contributed to the users being able to answer the questions correctly.

Training & Practice. Multiple communication provides maximum results in a short time but at the cost of losing some of the content. Some of the users reported that it was not that easy to comprehend both voices together, and therefore, retention of content in memory felt relatively lower than the sequential information. However, some of the users were of the view that practicing on such a system can improve comprehension in such systems as they reported that in the 1st study it was unexpected behavior for them to listen to two streams concurrently, but later brain got used to of it and started to process both the competing streams with less challenge.

Preference. Furthermore, to a pertinent question, 'whether you would prefer multiple information communication concurrently over the sequential flow of information?', the equal number of users answered 'Yes,' 'No,' and 'Maybe'. Many of the users who opted 'No' argued that in the concurrent form of delivery, they might miss a significant amount of information that could be a big problem when the information requirement is crucial and requires listening to it carefully and uninterruptedly. Therefore, users asked to provide them the authority regarding player controls to decide themselves whether they want two voices streams to be played concurrently or not. Those who opted 'may be' also argued that it would depend on the information-seeking context.

3 DISCUSSION

Contrary to our expectations, most of the participants found interval-based communication not very much helpful in comprehending multiple voice-based streams. The non-optimal utilization of auditory bandwidth and volume for auxiliary information could be the reasons. Therefore, the identification of optimal auditory bandwidth for the additional information stream can be a subject of an investigation to play the audio overlay.

Furthermore, the dichotic listening based on audio panning helped users in segregating both the streams from each other. Almost, all the users agreed that 'panning' was helpful in separating the content from each other. Therefore, another approach for investigation could be delivering continuous streams (study 1) dichotically anticipating that it would increase the comprehension.

Overall, the studies and response of the participants are encouraging to explore this avenue further. Although some of the

users manifested disinclination in multiple voice information delivery, their score in the questionnaire reveals that they performed well in comprehending the information from voice streams played concurrently. Many of the participants looked very keen to the development of new technologies that may assist them in meeting the daily challenges of extensive information.

3.1 Vinfomize Framework

Based on the feedback received from both the sighted and the visually challenged users, we have introduced a framework, called Vinfomize (V = Voice-based, info = Information, mize = optimization), for communicating multiple information concurrently. Many of the users opined that it depends on the context, whether they would listen to multiple voice-based information concurrently or go with the sequential form of communication. For example, the students with short time to prepare for the exams may opt the concurrent form of communication to revise the key concepts that they know already. Or, if the students do not have any time constraint, they might prefer to listen to the lectures sequentially for broad learning and understanding. It implies that the context is important, and control should rest with the users to opt the information presentation based on their information seeking context.

In our studies, we found that multiple information communication depends on the following factors:

- User auditory perception capabilities
- User information needs
- Information type
- Time constraint
- Physical context of the user

In Vinfomize, as illustrated in Fig 3, the concept of presenting multiple audio players having individually associated set of controls on a single page is introduced. Using these controls users may play multiple players together, or they can play a single player at any time. According to [12], 2-3 concurrent streams renders better results for information scanning. Vinfomize enables users to set the audio controls according to their need and receive optimum information from the system.

Furthermore, each component mentioned in the Vinfomize framework is briefly discussed below:

3.1.1 Audio API. Audio API could be any source that provides various types of audio streams/files that user may fetch to listen.

3.1.2 Playlists. Playlists are the organized lists that users may use to categorize the different type of content. Playlists can of various types, e.g., news, features, talk shows, songs, sports commentary, drama, etc.

3.1.3 Audio Players. As mentioned already, a page would have more than one audio players. Therefore, a user may play a playlist using an audio player. Users simultaneously can play more than one playlists using the equal number of audio players with the help of controls sets associated individually with each player. For example, a user may play news on audio player-1, songs on the audio player-2 and live commentary on the audio player-3 by applying suitable settings on each of the player using audio controls.

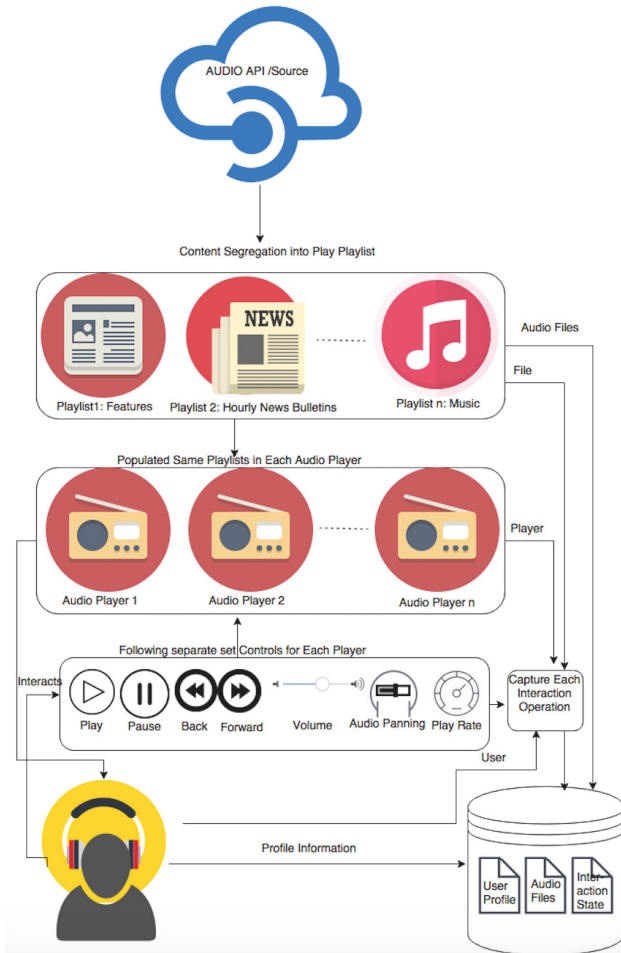


Figure 3: Vinformize: A framework for optimizing multiple voice-based information communication.

3.1.4 *Set of Controls.* Each of the control from the control panel that we used in Vinformize is explained below:

Play / Pause. A control button is provided to play or pause the audio stream on the relevant player. Interacting with this control on one player won't affect the other players.

Next / Previous Content. Similarly, forward and back buttons are provided in Vinformize using which a user may jump to the next track in the playlist or back to play a track. Besides these buttons, a seek bar is also provided that users may use to jump to the particular content within the audio stream.

Volume. As many of the participants had asked for this control because they believed that volume control would help them to segregate multiple streams, therefore, volume control is provided in Vinformize. Using the volume control users may set the sound pressure level of the associated player. Setting different sound pressure

level for different players may help the users better to segregate the multiple streams from each other.

Panning. This is a new control that we provided to induce the spatial difference between the competing streams.

This is a new control that we provided to induce the spatial difference between the competing streams. In our study, spatial difference or dichotic listening appeared one of the most important factors for segregating the competing streams. This panning control would help users to identify the source of the stream that eventually would help users to discriminate the streams from each other. Setting the extreme values on panning control would enable users to achieve dichotic listening. For example, for directing the two audio streams to separate ears, users may set pan control value to 1 for one stream and -1 for other. This setting would lead to playing the two streams together in the separate ears, first one in the right and ear, and the second one in the left ear.

Moreover, the panning can easily be extended to the n number of streams. For example, for 3rd stream pan value can be set to 0 which would give diotic effect to the users that means 3rd stream would be played in both the ears and this effect would give an impression that stream is being played over the head while the rest of the two streams are coming to the left and the right ear separately. Similarly, users may set any value from 1 to -1, and the values set by the users would determine the angle of the presentation of the stream to the users.

Playback-Rate. We also provided playback-rate control. Using this control a user can increase or decreases the audio playing speed. Playback-rate can help users to quickly listen to the information particularly in the sequential form of communication. The value of this control can be set from 0.5 to 2 to adjust the playing speed of the streams. In our experiment, as many users had also commented that the faster speed of information delivery hindered them to grasp the information, therefore, using this control, users would be able to set the play-rate according to their preference.

Users. The most important part of our framework is the user. Though this framework is applicable on system speakers, it is recommended that user should use earphones as it would not only provide a greater panning effect but would also minimize the masking.

3.1.5 *Interaction & Database.* Since communicating the concurrent information is in its developing or initial phase, therefore, we have incorporated a database in our framework to store all the users' interaction with such a system for post analysis by the research, and consequently, introduce new features and techniques for multiple information communication. Since this component is added in the framework for further research and enhancements, therefore, the designers in the industry may opt not to use this feature of the framework.

4 CONCLUSIONS & FUTURE WORK

In these studies, we aimed for communicating multiple voice-based information concurrently. In this investigation, users found the continuous form of delivery more appropriate than the interval-based method. However, they experienced dichotic audio technique

helpful in segregating multiple voice streams from each other. The spatial difference attribute with varying angles needs to be explored more along with other audio controls.

Moreover, based on the feedback received from the participants in this investigation, we have also introduced a framework to support the multiple information communication by employing multiple audio players associated with different control features. We are looking forward to developing a web-based prototype based on Vinfomize framework. It is expected that a large number of users with distinctive profiles from different walks of life would use our system. We would store all the interactions of the user as mentioned above and perform a comprehensive analysis of user interactions with the system. This analysis would expectedly give us further insight that how users interact with the multiple information communication systems, and consequently, would enable us to introduce new techniques for efficient information communication.

REFERENCES

- [1] Albert S Bregman. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.
- [2] Roxanne L Canosa. 2009. Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception (TAP)* 6, 2 (2009), 11.
- [3] Simon Carille and Daviid Schonstein. 2006. Frequency bandwidth and multi-talker environments. In *Audio Engineering Society Convention 120*. Audio Engineering Society.
- [4] E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.
- [5] Karen Church, Mauro Cherubini, and Nuria Oliver. 2014. A Large-scale Study of Daily Information Needs Captured in Situ. *ACM Trans. Comput.-Hum. Interact.* 21, 2, Article 10 (Feb. 2014), 46 pages. <https://doi.org/10.1145/2552193>
- [6] Mickael LD Deroche and John F Culling. 2011. Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *The Journal of the Acoustical Society of America* 130, 5 (2011), 2855–2865.
- [7] Mickael LD Deroche, John F Culling, Monita Chatterjee, and Charles J Limb. 2014. Roles of the target and masker fundamental frequencies in voice segregation. *The Journal of the Acoustical Society of America* 136, 3 (2014), 1225–1236.
- [8] Alan Dix, Janet E Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-Computer Interaction*. (2003).
- [9] Valérie Duthoit, Eric Enregle, Jean-Marc Sieffermann, Camille Michon, and David Blumenthal. 2017. Subjective contribution of vibrotactile modality in addition to or instead of auditory modality for takeover notification in an autonomous vehicle. In *Tenth International Conference on Advances in Computer-Human Interactions (ACHI2017)*. IARIA XPS Press, np.
- [10] Geo-News. 2016. Geo-News. <http://www.geo.tv>. [Online; accessed 01-May-2016].
- [11] Torkan Gholamalizadeh, Hossein Pourghaemi, Ahmad Mhaish, Gökhan Ince, and Damien Jade Duff. 2017. Sonification of 3D Object Shape for Sensory Substitution: An Empirical Exploration.
- [12] João Guerreiro. 2013. Using simultaneous audio sources to speed-up blind people's web scanning. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM, 8.
- [13] João Guerreiro and Daniel Gonçalves. 2015. Faster Text-to-Speeches: Enhancing Blind People's Information Scanning with Faster Concurrent Speech. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 3–11.
- [14] Antje Ihlefeld and Barbara Shinn-Cunningham. 2008. Spatial release from energetic and informational masking in a selective speech identification task. *The Journal of the Acoustical Society of America* 123, 6 (2008), 4369–4379.
- [15] Michael Johnston and Amanda J Stent. 2010. EPG: speech access to program guides for people with disabilities. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 257–258.
- [16] Kimron L Shapiro, Judy Caldwell, and Robyn E Sorensen. 1997. Personal names and the attentional blink: a visual "cocktail party" effect. *Journal of Experimental Psychology: Human Perception and Performance* 23, 2 (1997), 504.
- [17] Fu-Shing Sun, Xueying Kong, and Yi-Hua Weng. 2018. An Interactive Web System for Group Project Management and Peer Evaluation. In *Proceedings of the 2nd International Conference on Information System and Data Mining*. ACM, 103–106.

Appendix J

Publication 4 [Submitted]

——, “Investigating Efficient Speech-based Information Communication - A Comparison between the High-rate and the Concurrent Playback Designs,” *Journal on Multimodal User Interfaces (JMUI)*, vol. -, no. -, pp. 1–8, 2019, submitted

Investigating Efficient Speech-based Information Communication:

A Comparison between the High-rate and the Concurrent Playback Designs

First Author · Second Author · Third Author

Received: date / Accepted: date

Abstract This research aims to assist users to seek information efficiently while interacting with speech-based information, and reports on an experiment that tested two speech-based designs for communicating multiple speech-based information streams efficiently. In this experiment, a high-rate playback design and a concurrent playback design are investigated. In the high-rate playback design two speech-based information streams were communicated by doubling the normal playback-rate, and in the concurrent playback design two speech-based information streams were played concurrently. Comprehension of content in both the designs was also compared with the benchmark set from regular baseline condition. The results showed that the users' comprehension regarding the main information dropped significantly in the high-rate playback and the concurrent playback designs compared to the baseline condition. However, in answering the questions set from the detailed information, the comprehension was not significantly different in all three designs. It is expected that such efficient communication methods would foster, promote, improve, and increase user experience and productivity by providing information efficiently while interacting with an interactive system.

Keywords Speech Interface · Auditory Display · High-rate Playback · Concurrent Playback · Speech-based Information Communication · Comprehension

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

1 Introduction and Background

Multimodal User Interfaces enable users to perform their computing activities more freely and naturally as these interfaces provide users with multiple modes to interact with the system. These interfaces also support users to perform digital activities ubiquitously while doing other routine activities about daily life. In some interaction scenarios, while performing various activities simultaneously, users sometimes do not feel comfortable in interacting with computing devices using a visual interface. In such situations, an auditory display can enhance user experience positively [20].

The field of the auditory display is growing in maturity, and a large variety of techniques for conveying information through sound have been proposed by various researchers [3, 4, 8, 9]. In a non-speech-based auditory display, different techniques can be used to either enhance the visual display or communicate information using audio. Since the primary aim of auditory displays is to communicate information through sound, the use of speech seems reasonable, as humans in their daily life interact with each other using the same method that provides enormous flexibility and precision to exchange information [21]. In many ways, speech can appear to be the ideal solution for auditory displays. Presently, most of the speech interfaces communicate content through a single stream of speech.

Sequential speech interfaces, for example, digital audio streams, screen readers and voice messaging communicate speech-based information in a single speech stream; whereas, the users are capable of obtaining information from multiple sources concurrently [?, 26] and as well as on a high-rate playback [7]. Kramer [24] pointed out, speech has a low information transmission rate for continuously changing variables relative to the

bandwidth of the human auditory system. This mismatch implies that the sequential approach on regular-rate playback is under-utilising human perception capabilities and restricting users to perform optimally. There is a need to carry out research and develop design strategies in which information, such as speech, could be communicated efficiently to the user through concurrent multiple channels or with sequential high-rate playback.

Some researchers have proposed to address this challenge using high-playback speech rates, which use different forms of temporal compression to allow users to skim through spoken content at different rates [6,7,35]. Popular streaming platforms like YouTube, Udacity, and edX provide users with an option to set playback-rate according to their needs [30]. A lot of research is being carried out, and various models and techniques have been discussed by researchers to optimise playback speed of digital content [6,7,35].

The high-rate playback is one of the fast methods to communicate information quickly. The other productive quick design could be, concurrently playing two information streams on different topics to the users with regular playback-rate, as users are capable of listening and comprehending multiple information concurrently through their auditory perception [2]. Many researchers [13, 15–17, 19, 21, 22, 27, 29, 31–34, 37] have studied introducing concurrent communication and reported remarkable performance by participants in listening to two simultaneous messages, showing that a listener can process secondary information present in messages outside of immediate focus.

Schmandt and Mullins [31] introduced ‘AudioStreamer’, a tool that exploits people’s ability to separate two streams into distinct sources for effective browsing from multiple concurrent streams of real-time or stored audio. Hinde [21] explored how auditory displays can offer an alternative method for television experiences that depend on users’ desire to being able to attend to screen-based information visually. The results showed that offering sound-based secondary content from a smartphone after removing speech from a television program was the best auditory approach. For improving pilots’ situational awareness about the frequent changing state of systems information, Towers’ research [34] supported the use of spatial auditory displays within flight decks, showing that the use of concurrent spatial sonifications helped pilots to fly the aircraft more precisely. Guerreiro [19] conducted experiments with 30 visually impaired participants to compare the use of faster speech rates against the use of concurrent speech in the context of screen readers to ‘scan and find relevant information’. The results of this study [19] showed that concurrent

information streams with slightly faster playback-rate produce significantly faster ‘scanning’ for finding relevant content.

The researchers in the U.S. Naval Research Laboratory (NRL) for improving the Navy watch standing operations conducted a study [5] aimed at developing a set of comparative measures of attention and comprehension in a variety of multi-talkers information contexts involving concurrent and serial speech communications. In this research [5], authors by involving twelve (3 female, 9 male) participants from NRL, compared spatialized concurrent designs with sequential designs playing information at 75% faster playback-rate. In this research, four conditions were tested where listeners respectively heard two and four concurrent talkers and four sequential talkers (i.e., one at a time) speaking normally and 75% faster. In concurrent design, the spatial difference between the talkers was induced on the finding that spatial difference between the competing voice provides easy segregation between the voice streams [9,10,13,25]. In the experimental method, users were asked to perform two response tasks 1) identify the target noun phrases while listening, and 2) identify the relevant sentence (verbatim or semantically equivalent) to the spoken content immediately after listening. The first task aimed at determining the attentional abilities and the second task measured comprehension. However, both measures didn’t assess the level of comprehension depth in both concurrent and sequential approaches. The results showed that performance in the faster sequential condition was substantially higher than in either of the concurrent conditions. The study argues that sequential monitoring at synthetically faster rates of speech deserves further exploration as a possible alternative to concurrent monitoring [5].

In our study, we also compared the high-playback rate and the concurrent playback approaches by testing them with thirty-four (14 female and 20 male) general participants. We tested the concurrent design without involving the spatial difference between the competing information streams and compared it with the sequential design where speech playback-rate was doubled (100% faster). In our experiment, we determined the comprehension depth by comparing comprehension performance across several different formats of questions (main/detailed, implied/stated). We designed the questionnaire following the assessment pattern adopted in the standardised Discourse Comprehension Test (DCT) [23,28,36], which systematically assess the comprehension and retention of spoken narrative discourse by adults. The experiment also determined comprehension drop in both the quick designs in comparison to the regular sequential information communication.

2 Aims & Motivation

2.1 Aims

The fundamental aim of this study is to investigate the possibilities of communicating multiple-speech based information streams efficiently. The analysis in this chapter set out to satisfy the following questions: a) How different the comprehension appears for concurrent playback and the high-rate playback designs when compared to the baseline condition? b) Do both the high-playback rate and the concurrent playback design render similar comprehension? c) Does the comprehension pattern in all these designs remain the same?

2.2 Motivation:

The motivation of this experiment is to provide users with efficient methods enabling them to fulfil information needs quickly and efficiently while interacting with interactive systems. Based on the results of this study, an appropriate information communication approach can be adopted in speech-based interaction to communicate more information to listeners in an efficient manner. This study can also help to guide designs of sophisticated and information-heavy speech interaction methods. Quick communication of multiple information streams enhancing user experiences and understanding of speech-based information can help in listening to digital streams, screen reading, relevance scanning, scanning for specific information, notifications using a secondary audio channel, navigation, etc. [18].

3 Method

The experiment investigating the above three aims is outlined below.

3.1 Participants

After receiving the institutional Human Research Ethics Committee approval for the research protocol, user participation campaigns were launched that included: contacting people personally, sending emails to research communities, and pasting call for participation posters in common areas of the University. The participants were selected based on two criteria: 1) not having a significant hearing impairment, and 2) having competent English language skills. The second criterion was added because the listening experiment's content was in the English language. In total, 34 participants, 14 female,

and 20 male took part in the experiment after providing consent. The mean age of the participants was 26, with a standard deviation of 6.

3.2 Designs

Three designs were tested in this experiment that included the baseline, the high-rate playback, and the concurrent playback designs. Each of the design is explained individually, and also illustrated in Fig. 1 for further clarity.

3.2.1 Baseline

Under this condition, the baseline stimulus representing conventional speech-based communication was designed where the information stream in a female voice followed by the information stream in a male voice was presented sequentially without involving any auditory cue (see Fig. 1-b). Both the streams discussed different topics. The purpose of this design was to determine the comprehension benchmark based on users' comprehension in the baseline condition that could subsequently be used to evaluate users' comprehension in the high-rate playback and the concurrent playback designs.

3.2.2 High-rate Playback

In the high-rate playback condition, information streams were played following the baseline stimulus method with an only difference in playback-rate that was doubled (2x) from the normal recorded playback speed of the streams (see Fig. 1-c). An open-source Audacity software [1] was used for increasing the playback-rate that preserved the pitch, and other informational characteristics of the recorded speech. The purpose of this design was to test it for being able to quickly communicate multiple information in half of the time that is required in the baseline condition.

3.2.3 Concurrent Playback

Within the concurrent playback condition, the stimulus design was devised to communicate two speech-based information streams on separate topics concurrently (see Fig. 1-a). One stream was in a female voice, and the other was in a male voice. Again, the purpose of this design was to communicate two streams in half of the time that is required in the baseline condition.

In the concurrent playback condition, we had also created 11 more design variations that were tested in this experiment [13], but not discussed here because of the set scope of this paper.

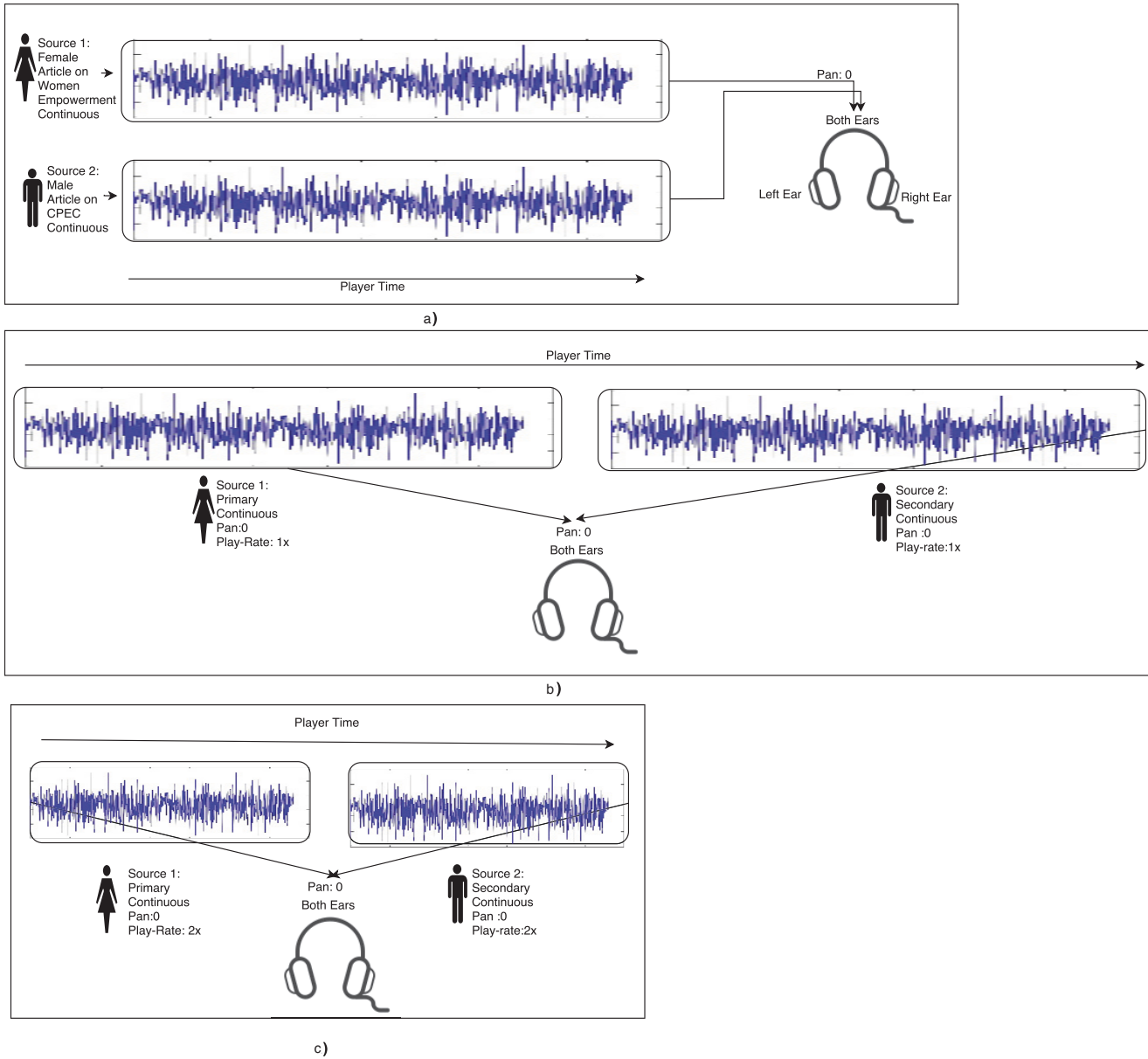


Fig. 1 Stimuli Designs: a) Concurrent: a stream in a female voice with a stream in a male voice, b) Baseline: a stream in a female voice followed by a stream in a male voice, c) high-rate playback: a stream in a female voice followed by a stream in a male voice with doubled playback-rate

3.3 Material: International English Language Testing System (IELTS)

For speech-based stimuli designs, International English Language Testing System (IELTS) was used [11]. The IELTS usually assesses the English language proficiency of non-native English speakers who want to study or work in a country where English is a primary language. The listening component of the IELTS evaluates the English listening abilities of users. The IELTS listening material was used in the experiment because it was

readily available and provided heterogeneous content in stereo-channelled audio files. For the experiment, twelve audio files containing monologue content having the sample rate of 44.1KHz and the bit rate of 16 were selected. In selection, six files were in female voices, and the remaining six were in male voices. From each monologue audio file, initial 58-70 seconds of meaningful content was extracted.

3.4 Stimuli Information

In total, 12 Continuous speech-based streams were obtained and processed following the design conditions mentioned above. The length of each rendered stimulus was within 58-70 seconds except the baseline design. The stimulus for the concurrent playback design was selected randomly.

3.5 Measures

After listening to each stimulus design, participants answered the questions, discussed in section 3.6. Since each stimulus was a combination of two streams and each stream had a set of 8 questions, therefore, each user was required to answer 48 questions having yes, no, and do not know options. A user's comprehension was measured based on the number of correct answers after listening to each stimulus.

In the previous experiments [13–15], users often pointed out that they did not know the answer and wanted to select a 'Do not know' option, which was not present, and therefore they were compelled to choose either 'Yes' or 'No'. In this situation, however, selecting 'Yes' or 'No' could lead to a correct, or incorrect response purely by chance, even where the users were not sure they were able to answer the question. This necessarily could have resulted in less accurate estimations of the user's comprehension of the stimulus content, with the assumption being that these participants will naturally choose one of the remaining two options equally. Therefore, in this experimental protocol the third option, 'Do not know' was included, in addition to the usual 'Yes' and 'No' options.

3.6 Questionnaire

For IELTS the new set of questions were prepared following the assessment pattern adopted in the standardised Discourse Comprehension Test (DCT) [23, 28, 36] that systematically assess the comprehension and retention of spoken narrative discourse by adults. For each stream, eight questions having yes, no, do not know options as answers were created. The questions were arranged in assessment categories to assess the level of comprehension by the users. For each following category, two questions were arranged:

- Main Information Stated (MIS)
- Main Information Implied (MII)
- Detailed Information Stated (DTS)
- Detailed Information Implied (DTI)

The questions in the MIS were constructed from the main stated information of the story. These questions assess how much a participant had comprehended the main idea that was repeated or elaborated by other information in the story (main information). The MII questions were based on the information that was not directly discussed in the story, but a user had to infer it from the stated main information. The questions in the DTS were framed from the stated information of the story that estimated the comprehension of detailed information. The detailed information was mentioned only once and was not elaborated by other information in the story. The DTI questions were based on the information that was not directly explained in the story, but the users had to infer from the detailed information. The implied questions examined whether a user was able to make a mental map or bridging assumptions of the information.

3.7 Apparatus

To minimize the participation time for completing the tests and convenience the web-based system was designed to play stimuli. The web system was developed using PHP, MySQL, JQuery, HTML5, CSS, and Bootstrap. The web system was accessible using the latest web browsers where different audio players each playing one stimulus design were presented on the screen along with the questions under the relevant stimulus player. Most of the tests were conducted in quiet purpose-built creativity and cognition studios (CCS) of the University of Technology, Sydney. Three identical i-Mac computers having 2.7GHz quad-core Intel Core i5 processor, 8 GB RAM, installed with Yosemite 10.10.5 OS were arranged in the studio. To listen to the audio stimuli Beyerdynamic's DT770 250 OHM headphones were used that were connected to the headphones jack of the computer. Since three computers were used in the studio, therefore, at a time, up to three participants could engage in the experiment simultaneously. The web-based system also enabled to involve users from around the world. Among 34, six users participated remotely from the United States, Pakistan, Saudi Arabia and Hungary by accessing our web-system using their devices.

3.8 General Procedure

The selected users were verbally briefed on the study protocol before the start of the experiment, and also the instructions were presented on the screen after registration. Before starting the experiment, the users entered their demographic profile information that included,

name, age, qualification, first language, country, hearing impairment and type of computer & headphones used in case of participating from outside of the CCS. All users' responses were stored in the MySQL database for the post-experiment analysis discussed below. The dataset generated during the current study are available at figshare [12].

4 Results

The following three aspects are covered in the result analysis:

- The proportion of users' selection of options for answering the questions is evaluated in sub-section 4.1.
- The users' performance in comprehending information in terms of answering questions correctly in all three designs, i.e., baseline, high-rate playback, and the concurrent playback designs is compared in sub-section 4.2.
- The depth of information comprehension in all three designs is assessed in sub-section 4.3.

4.1 Proportion Analysis

This section evaluates the proportion of selecting options for answering the questions in all three designs individually.

In the baseline design, as shown in figure 2, the users frequently selected 'Do not Know' option for both types of 'Yes' & 'No' correct answers. This showed that when the correct answer was 'No', 23% responses were selected 'Do not know' by the users. This percentage was significantly higher than the 8% 'Do not Know' responses in the condition when the correct answer was 'Yes'. Also, The percentage, 18%, of selecting 'Yes' as a wrong answer was higher than the 14% of selecting 'No' as a wrong answer.

In the higher-play-rate design, as shown in figure 2, the proportion of selecting 'Do not Know' remained higher than the baseline conditions. In this approach, when the correct answer was 'No', the proportion of selecting Do not know was 29% and when the correct answer was 'Yes' the percentage remained 25%. Moreover, contrary to the baseline condition, the percentage of selecting "Yes", 18%, as a wrong answer remained lesser than the percentage of selecting "No", 22%, as a wrong answer.

Regarding the concurrent playback design, figure 2 shows the proportion of response submitted by the users. Similar trends appeared in this design as seen in the high-rate playback design. In this design, when the correct answer was 'No', the proportion of selecting Do not know was 35% and when the correct answer was

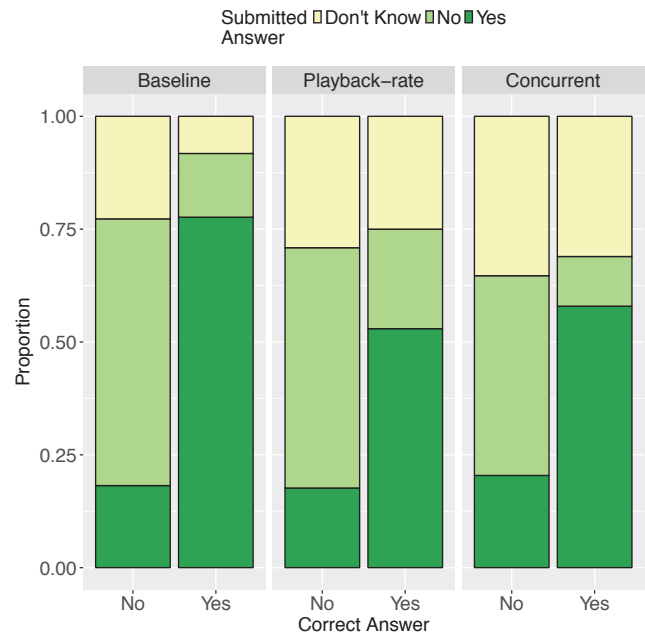


Fig. 2 The proportion of users responses.

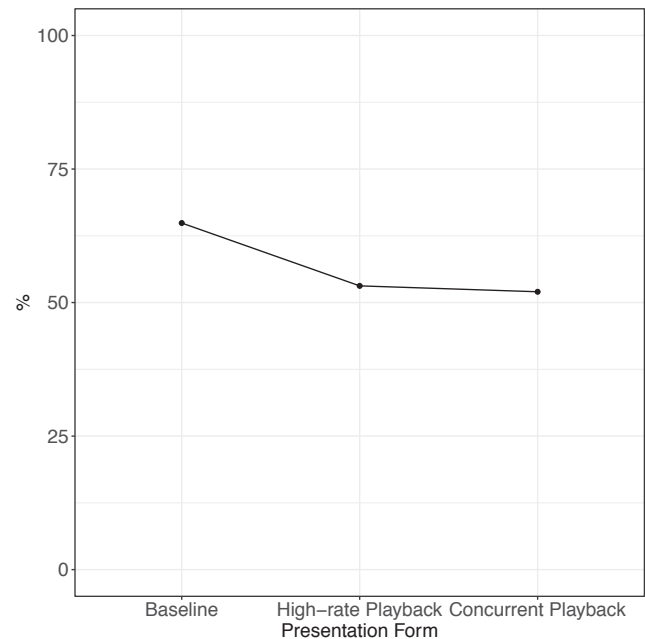


Fig. 3 The percentage of correct answers.

'Yes' the percentage remained 31%. The proportion of giving correct responses and selection of 'do not know' appeared similar as seen in the high-rate playback design.

Overall, in all three approaches, users selected all three 'Yes,' 'No' and 'Do not Know' options as the answers to the questions.

4.2 Comprehension Performance Analysis

In the second part of this analysis, the comprehension is assessed by calculating the percentage of correct answers for all three designs. For percentage calculation, the user's response matching to the expected answer was counted as the correct answer whereas the opposite answer or the selection of 'Do not Know' options were considered as the wrong answer. The assessments of all three designs are discussed individually.

In the baseline design analysis, the percentage of the correct answer was calculated to set a benchmark. As figure 3 shows, 63% questions were correctly answered by the users. Inversely, 37% questions either were answered incorrectly, or users did not know the answers. This implies that users could not fully understand the content to answer all the question correctly. The percentage, 63%, of giving the correct answer in the baseline sequential information communication set the benchmark to compare users' comprehension in the high-rate playback and the concurrent playback designs.

Regarding the high-rate playback design, figure 4 shows the percentage of correct answers. In this approach, the percentage of correct answers remained 53% that shows the users' comprehension performance remained significantly lower ($p < 0.001$) than the baseline design.

Similarly in the concurrent playback design, as shown in figure 4, the percentage of correct answers remained 52%. This indicates that the users' comprehension in the concurrent playback design remained similar ($p = 0.761$) to the high-rate playback design.

Overall, both the high-rate playback and the concurrent playback designs performed similarly in delivering multiple information. However, they remained significantly lower ($p < 0.001$) than the benchmark set from the baseline approach.

4.3 Comprehension Depth Analysis

The third part of the analysis started with an evaluation of the comprehension depth of the content in the baseline design and then followed by the high-rate playback and the concurrent playback designs.

In the baseline condition, the specific percentage concerning MIS, MII, DTS, and DTI is calculated to set a benchmark that was later used to compare the comprehension in both the quick designs. Figure 4 using the red line shows the analysis of the Baseline design. The percentage of correct answers to the questions set from the MIS remained 85% whereas, in the MII, DTS, and DTI, it remained 72%, 51%, and 51% respectively.

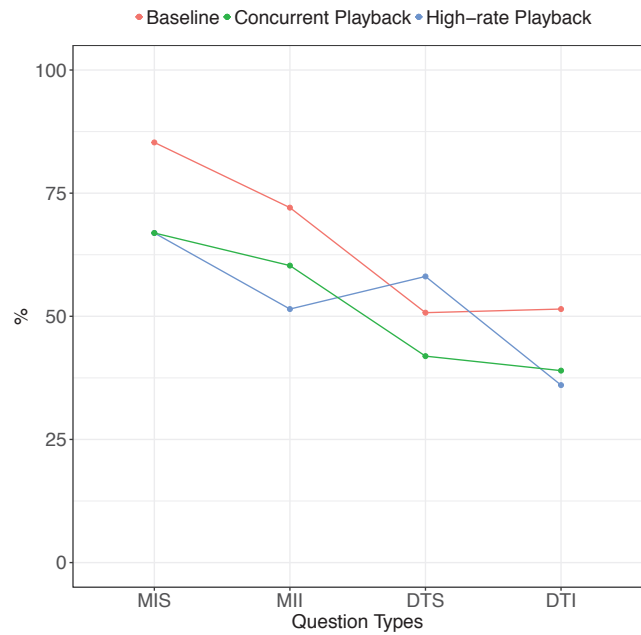


Fig. 4 Comprehension depth analysis.

The analysis shows that the comprehension of the MIS remained significantly higher compared to the other information categories.

Following the pattern adopted in the analysis of baseline design, the comprehension depth was evaluated for the high-rate playback design. Figure 4 using the blue line shows the percentage of correct answers in the concurrent playback design concerning the MIS, MII, DTS, and DTI. In this design, the percentage of correct answers to the questions set from the MIS remained 67% whereas in the MII, DTS, and DTI it remained 51%, 58%, and 36% respectively.

In the concurrent playback design, the percentage of correct answers to the questions set from the MIS remained 67% whereas in the MII, DTS, and DTI it remained 60%, 41%, and 39% respectively. This shows that the pattern of information comprehension in this concurrent playback approach remained similar to the comprehension depth calculated in the high-rate playback design, except one condition where the percentage regarding the DTS in the high-rate playback design remained higher than the concurrent playback design. The comprehension assessment for concurrent playback design is reflected using the green line in figure 4.

Besides calculating comprehension depth, each MIS, MII, DTS, DTI data point (percentage) in the high-rate playback design and the concurrent playback design are also statistically compared with the relevant data points in the benchmark set from the baseline condition and mentioned in Table 1. In almost all the data-points re-

lated to the main information, the comprehension in both the quick designs remained lower compared to the baseline condition, however, in the data-points related to detailed information, comprehension was not significantly different in all the three designs.

Table 1 Results of one-to-one proportion comparison between the correct answers in the baseline and both quick designs w.r.t MIS, MII, DTS, DTI, with Bonferroni correction, *** < 0.001

Design	MIS	MI	DTS	DTI
High-rate Playback	***	***	0.273	0.014
Concurrent Playback	***	0.054	0.181	0.051

5 Discussion

The users' comprehension regarding the main information dropped significantly in the high-rate playback and the concurrent playback designs compared to the baseline condition. However, in answering the questions set from the detailed category, the comprehension was not significantly different in all three designs. In both quick designs, users were able to answer more than 50% of the questions correctly. In the analysis of the results, both designs, the high-rate playback and the concurrent, performed similarly in communicating information as users' comprehension score was almost the same. User's ability to answer over 50% of questions shows that both approaches have the potential to be used for communicating multiple information streams quickly. Therefore, further investigations should be carried out to explore more possibilities and come up with optimal designs for efficient multiple information communication.

Additionally, as discussed in the methods section, the questions were arranged in 4 categories MIS, MII, DTS, DTI formed by information repetition in the content to assess the comprehension depth. It was expected that the users content comprehension would remain in the same order mentioned above, as the main information was repeated multiple times in the content whereas the detailed information was played once only in the content. In all three designs, users' comprehension was higher in MIS and MII and was lower in DTS and DTI, except for the comprehension in DTS in the high-rate playback design. This shows that in all three designs, the pattern of comprehension depth was similar.

The users selected all three 'Yes,' 'No' and 'Do not Know' options as answers to the questions. In all the speech-based designs a similar pattern of proportion for

answering the questions was seen. However, the proportion of giving correct responses was lower, and the selection of 'do not know' option was higher in quick designs compared to the baseline condition. This further shows that the users faced some comprehension challenges in both the quick designs.

Overall, this research work contributes to pursuing the 'design for all paradigm' [18] that aims to enable sighted users and as well the users having visual impairment to interact with digital information applications efficiently, e.g., a) Listening to information streams, b) Finding relevant information items, c) Scanning for specific information etc. This study compared two designs, one from each approach, to communicate multiple information efficiently. There can be many configurations in each approach that can be experimented for efficient communication of audio streams.

In comparing the viability of the efficient designs, both approaches render similar comprehensibility; however, there are some trade-offs. The concurrent communication comparing to the higher-rate playback is attentionally more demanding [5]. Researchers who consider faster rates of speech as an alternative to listening to multiple speech communications further explored the sequential speech at higher-playback rate [6,7,35]. However, the concurrent playback design has certain advantages over the high-rate playback design. Some of the main advantages of concurrent communication are:

- Supports live streaming - for example, if two radio programs are being broadcast live at the same time, users would be able to listen to both of them at the same time with concurrent presentation. In other words, concurrency may help users to listen to two different streams simultaneously.
- Selecting and attending an information stream - for example if two streams are provided concurrently users using their selection and attention abilities may switch focus immediately towards the information stream that carries high user interest.
- Divided attention - Users may get the gist from both the streams at the same time using divided attention.

These advantages show that in some of the contexts, the concurrent playback approach can provide more efficient communication for the multiple streams to the users. Therefore, we will continue exploring the concurrent playback approach, and analyse different designs to come up with an optimal design that could concurrently communicate multiple speech-based information streams while creating minimum cognitive load.

6 Limitations and Future Work

This detailed experiment was designed to comprehensively assess users' comprehension in quick speech-based communication designs that required users to engage with the experiment for the extended time. We minimised the required time by designing a usable interactive web system that brought users' participation to less than 45 minutes. Still, users reported a high cognitive load, which might have impacted users' comprehension. Since users' comprehension is reported comparatively, therefore, the results and findings given in this paper remain valid.

This study shows the potential of communicating multiple information using the high-rate playback approach and the concurrent playback approach. There can be many configurations in each approach that can be tested for efficient communication of audio streams. We will continue exploring the concurrent playback approach, and analyse different designs. It is expected that such novel efficient communication designs would foster, promote, improve, and increase user experience and productivity by providing information efficiently.

Acknowledgements This research was supported by the School of Software, Faculty of Engineering and IT, University of Technology Sydney, Australia.

References

1. Audacity: Audacity. <https://www.audacityteam.org/>. [Online; accessed 23-Dec-2018]
2. Bakker, S., van den Hoven, E., Eggen, B.: Knowing by ear: leveraging human attention abilities in interaction design. *Journal on Multimodal User Interfaces* **5**(3-4), 197–209 (2012)
3. Brazil, E., Fernström, M.: Investigating concurrent auditory icon recognition. In: *Proceedings of the 12th International Conference on Auditory Display (ICAD)*, pp. 51–58. Georgia Institute of Technology (2006)
4. Brazil, E., Fernstrom, M., Bowers, J.: Exploring concurrent auditory icon recognition. In: *Proceedings of the 15th International Conference on Auditory Display (ICAD)*, pp. 1–4. Georgia Institute of Technology (2009)
5. Brock, D., McClimens, B., Trafton, J.G., McCurry, M., Perzanowski, D.: Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech. In: *Proceedings of the 14th International Conference on Auditory Display (ICAD)*, pp. 1–8. Georgia Institute of Technology (2008)
6. Brock, D., Wasylyshyn, C., McClimens, B.: Word spotting in a multichannel virtual auditory display at normal and accelerated rates of speech. In: *Proceedings of the 22nd International Conference on Auditory Display (ICAD)*, pp. 130–135. Georgia Institute of Technology (2016)
7. Brock, D., Wasylyshyn, C., McClimens, B., Perzanowski, D.: Facilitating the watchstander's voice communications task in future navy operations. In: *MILCOM 2011 Military Communications Conference*, pp. 2222–2226 (2011). DOI 10.1109/MILCOM.2011.6127692
8. Brungart, D.S., Ericson, M., Simpson, B.D.: Design considerations for improving the effectiveness of multitalker speech displays. In: *Proceedings of the 2002 International Conference on Auditory Display (ICAD)*, pp. 1–7. Georgia Institute of Technology (2002)
9. Brungart, D.S., Simpson, B.D.: Distance-based speech segregation in near-field virtual audio displays. In: *Proceedings of the 2001 International Conference on Auditory Display (ICAD)*, pp. 169–174. Georgia Institute of Technology (2001)
10. Brungart, D.S., Simpson, B.D.: Optimizing the spatial configuration of a seven-talker speech display. *ACM Transactions on Applied Perception (TAP)* **2**(4), 430–436 (2005)
11. Council, B.: IELTS. <https://www.ielts.org> (2019). [Online; accessed 30-Apr-2019]
12. ul Fazal, M.A., Ferguson, S., Johnston, A.: Multiple Speech-based Information Communication (2019). DOI 10.6084/m9.figshare.8214776.v1. URL https://figshare.com/articles/Multiple_Speech_based_Information_Communication/8214776
13. Fazal, M.A.u., Ferguson, S., Karim, M.S., Johnston, A.: Concurrent Voice-Based Multiple Information Communication: A Study Report of Profile-Based Users' Interaction. In: *145th Convention of the Audio Engineering Society*. Audio Engineering Society (2018)
14. Fazal, M.A.u., Ferguson, S., Karim, M.S., Johnston, A.: Vinfomize: A Framework for Multiple Voice-based Information Communication. In: *Proceedings of the 3rd International Conference on Information System and Data Mining*, pp. 1–7. ACM - Accepted (2019). Accepted - to be published.
15. Fazal, M.A.u., Shuaib, K.: Multiple information communication in voice-based interaction. In: *Advances in Intelligent Systems and Computing*, pp. 101–111 (2017)
16. Feltham, F., Loke, L.: Felt sense through auditory display: A design case study into sound for somatic awareness while walking. In: *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, pp. 287–298. ACM (2017)
17. Ferguson, S., Cabrera, D.: Exploratory sound analysis: sonifying data about sound. pp. 1–8. *International Community for Auditory Display* (2008)
18. Guerreiro, J.: Towards screen readers with concurrent speech: where to go next? *SIGACCESS Accessibility and Computing* (114), 12–19 (2016)
19. Guerreiro, J., Goncalves, D.: Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing* **8**(1) (2016). DOI 10.1109/CVPR.2016.105
20. Hermann, T.: Taxonomy and definitions for sonification and auditory display. In: *Proceedings of the 14th International Conference on Auditory Display (ICAD)*, pp. 1–8. Georgia Institute of Technology (2008)
21. Hinde, A.F.: Concurrency in auditory displays for connected television. Ph.D. thesis, University of York (2016)
22. Ikei, Y., Yamazaki, H., Hirota, K., Hirose, M.: vCocktail: multiplexed-voice menu presentation method for wearable computers. In: *Virtual Reality Conference*, pp. 183–190. IEEE (2006). DOI 10.1109/VR.2006.141
23. Iyer, N., Thompson, E.R., Simpson, B.D., Brungart, D., Summers, V.: Exploring auditory gist: Comprehension of

- two dichotic, simultaneously presented stories. In: Proceedings of Meetings on Acoustics, vol. 19, pp. 050158–050158. Acoustical Society of America (2013). DOI 10.1121/1.4800507
24. Kramer, G.: An introduction to auditory display in kramer, g.(ed.) auditory display (1994)
 25. McGookin, D.K., Brewster, S.A.: An investigation into the identification of concurrently presented earcons. In: Proceedings of 2003 International Conference on Auditory Display (ICAD), pp. 42–46. Georgia Institute of Technology (2003)
 26. McGookin, D.K., Brewster, S.A.: Understanding concurrent earcons: applying auditory scene analysis principles to concurrent earcon recognition. *ACM Transactions on Applied Perception* **1**(2), 130–155 (2004). DOI 10.1145/1024083.1024087
 27. Mullins, A.T.: *Audiostreamer: Leveraging The Cocktail Party Effect for Efficient Listening*. Ph.D. thesis, Massachusetts Institute of Technology (1996)
 28. Obermeyer, J.A., Edmonds, L.A.: Attentive reading with constrained summarization adapted to address written discourse in people with mild aphasia. *American Journal of Speech-Language Pathology* **27**(1S), 392–405 (2018)
 29. Parente, P.: *Clique: Perceptually based, task oriented auditory display for GUI applications*. Ph.D. thesis, The University of North Carolina at Chapel Hill (2008)
 30. Patel, D., Ghosh, D., Zhao, S.: Teach Me Fast: How to Optimize Online Lecture Video Speeding for Learning in Less Time? In: Proceedings of the Sixth International Symposium of Chinese CHI, pp. 160–163. ACM (2018)
 31. Schmandt, C., Mullins, A.: *AudioStreamer: Exploiting simultaneity for listening*. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 218–219. ACM (1995). DOI 10.1145/223355.223533
 32. Song, H.J., et al.: Evaluation of the effects of spatial separation and timbral differences on the identifiability of features of concurrent auditory streams (2011)
 33. Tordini, F., Bregman, A.S., Cooperstock, J.R.: Prioritizing foreground selection of natural chirp sounds by tempo and spectral centroid. *Journal on Multimodal User Interfaces* **10**(3), 221–234 (2016)
 34. Towers, J.A.: *Enabling the Effective Application of Spatial Auditory Displays in Modern Flight Decks*. Ph.D. thesis, The University of Queensland (2016)
 35. Wasylshyn, C., McClimens, B., Brock, D.: Comprehension of speech presented at synthetically accelerated rates: Evaluating training and practice effects. In: Proceedings of the 16th International Conference on Auditory Display (ICAD), pp. 133–136. Georgia Institute of Technology (2010)
 36. Welland, R.J., Lubinski, R., Higginbotham, D.J.: Discourse Comprehension Test Performance of Elders With Dementia of the Alzheimer Type. *Journal of Speech Language and Hearing Research* **45**(6), 1175 (2002). DOI 10.1044/1092-4388(2002/095)
 37. Werner, S., Hauck, C., Roome, N., Hoover, C., Choates, D.: Can VoiceScapes assist in menu navigation? In: Proceedings of the Human Factors and Ergonomics Society, vol. 2015, pp. 1095–1099 (2015). DOI 10.1177/1541931215591157

Appendix K

Publication 5

M. A. u. Fazal, S. Ferguson, and A. Johnston, "Investigating Concurrent Speech-based Designs for Information Communication," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ACM. New York, NY, USA: ACM, 2018, pp. 1–8

Investigating Concurrent Speech-based Designs for Information Communication

Muhammad Abu ul Fazal
Creativity & Cognition Studios,
University of Technology, Sydney
Sydney, NSW, Australia
Muhammad.AbuUlFazal@uts.edu.au

Sam Ferguson
Creativity & Cognition Studios,
University of Technology, Sydney
NSW, Australia
Samuel.Ferguson@uts.edu.au

Andrew Johnston
Creativity & Cognition Studios,
University of Technology, Sydney
NSW, Australia
Andrew.Johnston@uts.edu.au

ABSTRACT

Speech-based information is usually communicated to users in a sequential manner, but users are capable of obtaining information from multiple voices concurrently. This fact implies that the sequential approach is possibly under-utilizing human perception capabilities to some extent and restricting users to perform optimally in an immersive environment. This paper reports on an experiment that aimed to test different speech-based designs for concurrent information communication. Two audio streams from two types of content were played concurrently to 34 users, in both a continuous or intermittent form, with the manipulation of a variety of spatial configurations (i.e. Diotic, Diotic-Monotic, and Dichotic). In total, 12 concurrent speech-based design configurations were tested with each user. The results showed that the concurrent speech-based information designs involving intermittent form and the spatial difference in information streams produce comprehensibility equal to the level achieved in sequential information communication.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI); Empirical studies in HCI;*

KEYWORDS

Concurrent audio, Speech-based Information Comprehension, Dichotic listening, Diotic listening, Information Comprehension Study, Spatial cues, Intermittent & continuous audio presentation

ACM Reference Format:

Muhammad Abu ul Fazal, Sam Ferguson, and Andrew Johnston. 2018. Investigating Concurrent Speech-based Designs for Information Communication. In *Audio Mostly 2018: Sound in Immersion and Emotion (AM'18)*, September 12–14, 2018, Wrexham, United Kingdom. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3243274.3243284>

1 INTRODUCTION & BACKGROUND

Voice-based interaction enables users to interact with a computer system in an immersive environment [17]. Conventional voice-based interaction designs provide communication in a sequential

form to the user, which theoretically underutilizes natural human perception capabilities [8, 9], as the single channel acts as a bottleneck for the amount of information that can be communicated in a given time [23]. User's auditory channel facilitates to concurrently listen or get a gist of the information from multiple sources [10] within their surroundings. Since users are capable of listening and comprehending multiple streams of information concurrently through auditory perception, and also have growing information needs [6], therefore, one possible solution to maximise information communication is to present multiple streams of speech-based information concurrently.

Besides studies to communicate multiple information concurrently in non-speech audio [4, 24, 25], recent studies on speech-based natural voices have also reported remarkable performance by participants in listening to two concurrent messages. These psychological studies have established that listeners can process information present in messages outside of the immediate auditory focus [7, 18–20, 22]. A listener can selectively read out 'secondary' information from working memory after the message ends [5, 7]. This selective readout from parallel sources is facilitated by various voice signal cues and users' personal and contextual circumstances [2, 3].

In the voice-based interaction concerning a computer system, Fazal and Karim explored the design approach to communicate two speech-based voices concurrently through an empirical study [11]. The results validated that multiple information communication is possible using voice in Human-machine interaction. Users were able to discriminate the voice and using their selection and attention abilities they were able to get multiple information meaningfully in lesser time. Importantly, users showed interest in concurrent multiple information communication. Similarly, Iyer et al. carried out experiments to understand the amount of the information comprehension, and also the nature of the semantic processing in concurrent information communication [16]. The results of these experiments identified that the participants were able to apprehend the main idea of the *unattended* story to a level that was higher than chance. Guerreiro and Gonçalves conducted experiments with sighted and with visually impaired persons to determine people's ability to find important speech content from two, three, or four speech channels played concurrently [12, 14, 15]. The study established: (1) Both the sighted and visually impaired users can successfully scan information from concurrent speech-based streams with no significant difference in their performance. (2) Two concurrent voices appeared to result in higher performance than three, and even further, four. (3) The spatial difference in sources appeared the best cue in concurrent speech. .

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AM'18, September 12–14, 2018, Wrexham, United Kingdom

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-6609-0/18/09...\$15.00

<https://doi.org/10.1145/3243274.3243284>

This paper is arranged as follows. The aims & motivation followed by a method of the experiment are presented. An analysis of the results of experiment and discussion follow.

2 AIMS & MOTIVATION

2.1 Aims

This study aims to examine designs for speech communication that can communicate concurrent speech-based information equal to the information transfer efficiency that is achieved in conventional sequential speech-based information communication.

2.2 Motivation

If this study is successful, concurrent speech-based communication designs that render better information communication can be adopted in speech-based interaction to communicate more information to listeners in an efficient manner and can help to guide the design of complex and information-heavy speech interaction methods. The concurrent speech can be helpful in listening to two TV streams, relevance scanning, scanning for specific information, notifications using a secondary audio channel, TV navigation and subtitles, assisted navigation etc. [13].

3 METHOD

The standardized experiment discussed in this paper is an extension of the work carried out by Fazal and Karim, Iyer et al., and Guerreiro and Gonçalves. The method adopted for this experiment is outlined below.

3.1 Participants

After receiving institutional Human Research Ethics Committee approval for the research protocol, user participation campaigns were launched. The participants were selected based on two criteria: 1) not having a significant hearing impairment, and 2) having competent English language skills, as the listening experiment's content was in the English language. In total, 34 participants, 14 high pitch and 20 low-pitched, took part in the experiment after providing consent. The mean age of the participants was 26 with the standard deviation of 6.

3.2 Design

3.2.1 Concurrent Condition. Within the concurrent condition, initially, six distinct stimuli designs were devised to communicate two speech-based information streams on separate topics concurrently. One stream was in the high pitched (female) voice, and the other was in the low pitched (male) voice. From six, three designs followed the first form, and the remaining three followed the second form of communication from the following list:

- Continuous High-Pitched Stream with Continuous Low-Pitched Stream (Continuous)
- Continuous High-Pitched Stream with intermittent Low-Pitched Stream (Intermittent).

Each of the continuous and intermittent based stimuli design was individually applied with one of the following three pan conditions to involve a spatial difference in streams presenting streams to the specific ear(s):

- 0,0 – Diotic (Both Streams in **both ears**)
- 0,100 – Diotic-Monotic (High-Pitched stream in **both ears** whereas the Low-Pitched stream in the **right ear**)
- -100,100 – Dichotic (High-Pitched stream in the **left ear** whereas the Low-Pitched voice stream in the **right ear**)

All the six design methods were repeated on two types of audio content material that increased concurrent stimuli designs to 12. The audio types of content material were:

- Discourse Comprehension Test (DCT)
- International English Language Testing System (IELTS)

Each of the rendered concurrent stimuli design is described in Table 1 and illustrated in Figure 1 for further clarity.

3.2.2 Baseline Condition. Under this condition, a baseline stimulus representing the conventional speech-based communication was designed where the continuous high pitched information stream followed by a continuous low-pitched information stream was presented sequentially without involving spatial difference. The purpose of this design was to determine a benchmark of user comprehension in the baseline condition that could subsequently be used to evaluate users' comprehension in concurrent condition.

Besides baseline stimulus, another sequential stimulus (Seq-2x) was also designed where streams were played following the baseline stimulus method with the only difference in play-rate that was doubled (2x). The purpose of this design was to test another design to communicate multiple information in unit time as shown in Fig. 1-d. This design is not discussed in the result and discussion sections to maintain the simplicity by limiting the scope of this paper.

3.3 Material

For speech-based stimuli designs, two types of content resources were used:

3.3.1 Discourse Comprehension Test (DCT). The commercially available Discourse Comprehension Test (DCT) [16, 21, 26] is a standardized test to primarily assess the comprehension and retention of spoken narrative discourse by adults suffering from aphasia. The test contains 12 stories where each story having a length from 73 to 95 seconds describes a humorous situation. The material purchased from [1] was received on a CD having twelve mono-channelled audio tracks each presenting a story in the male (low-pitched) voice. To use the stories in the experiment, each track was exported into .wav format with the sample rate of 44.1KHz and the bit rate of 16 using the Apple iTunes software. Since the conceived stimuli designs were to be discriminable by pitch (gender) i.e. high-pitched (female) voice and low-pitched (male) voice, therefore, the pitch of the six from twelve tracks was changed by increasing it 17% from the default low frequency male voice using Audacity software (using Sound-touch). This increase in pitch transformed the male (low-pitched) voice into a female (high-pitched) voice. Resultantly, it converted six stories in a high-pitched voice, and six in a low-pitched voice.

3.3.2 International English Language Testing System (IELTS). The IELTS listening material was also used in the experiment because it was readily available and provided heterogeneous content

Table 1: Speech-based Concurrent Communication Designs

Concurrent Design			Primary Stream			Secondary Stream		
Content Type	Form	Pan Condition	Voice	Presentation	Ear	Voice	Presentation	Ear
DCT	Continuous	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Both
DCT	Continuous	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Right
DCT	Continuous	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Continuous	Right
DCT	Intermittent	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Both
DCT	Intermittent	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Right
DCT	Intermittent	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Intermittent	Right
IELTS	Continuous	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Both
IELTS	Continuous	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Right
IELTS	Continuous	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Continuous	Right
IELTS	Intermittent	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Both
IELTS	Intermittent	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Right
IELTS	Intermittent	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Intermittent	Right

in stereo-channelled audio files. For the experiment, 12 audio files containing monologue content having the sample rate of 44.1KHz and the bit rate of 16 were selected. In selection, six files were in the male (low-pitched) voice and remaining six were the female (high-pitched) voice. From each monologue file, initial 58-70 seconds of the meaningful content was extracted.

3.4 Stimuli Information

In total, 24 continuous speech-based streams were obtained and processed from both types of material. For having the intermittent streams, the contents of the half of the continuous stream in low-pitched voice were broken into chunks by giving silent intervals of 5 to 10 seconds in them. Each stream was repeatedly applied with each of the three pan conditions 0, 100 and -100 that rendered 72 (24 x 3) streams where 36 were in the high-pitched voice, and 36 (18 continuous and 18 intermittent) were in the low-pitched voice. Then each of the rendered low-pitched stream was repeatedly combined with the high-pitched stream of the same material using the Audacity software for Mac. This multiplication generated 216 combinations to incorporate randomization in the experiment for minimizing the combinational effect in the analysis. From 216 stimuli, randomly 12 (6 DCT + 6 IELTS) were presented to each user where each stimulus was a representation of one of the designs mentioned in table 1. The length of each rendered stimulus was within 55 to 90 seconds except the baseline design. Besides the 12 concurrent designs, the additional two designs, baseline and Seq-2x, were presented to the participants.

3.5 Measures

After listening to each stimulus design, participants answered the questions, discussed in section 3.6, from the stimuli. Since each stimulus was the combination of two streams and each stream had a set of 8 questions, therefore, a user answered 224 questions having yes/no/don't options. The user comprehension was measured on the basis of the number of giving correct answers after listening to each stimulus.

In previous experiments by the authors [11], users often pointed out that they did not know the answer and were looking to select a

'Don't know' option, which wasn't present, and therefore were compelled to choose either 'Yes' or 'No'. This necessarily has resulted in less accurate estimations of user comprehension of the stimulus content, with the assumption being that these participants will naturally choose one of the remaining two options equally. Therefore, in this experimental protocol a third option, 'Don't know' was included, in addition to the usual 'Yes' and 'No' user responses.

3.6 Questionnaire

The DCT material was accompanied with the default questions that were used in the experiment as is, however, for IELTS new questions following the DCT pattern were prepared. Each story had eight questions having yes/no/don't know answers. The questions were arranged in assessment categories to assess the depth of comprehension by the users. For each following category type, two questions were arranged:

- Main Information Stated (MIS)
- Main Information Implied (MII)
- Detailed Information Stated (DTS)
- Detailed Information Implied (DTI)

The questions in MIS were constructed from the main stated information of the story. These questions assessed how much a participant had comprehended the main idea that was repeated or elaborated by other information in the story (main information). The MII questions were based on the information that was not directly discussed in the story, but a user had to infer it from the stated main information. The questions in DTS were framed from the stated information of the story that estimated the comprehension of detailed information. Detailed information was mentioned only once and not elaborated by other information in the story. DTI questions were based on the information that was not directly explained in the story, but a user had to infer from the detailed information. The implied questions examined whether a user was able to make a mental map or bridging assumptions of the information or not.

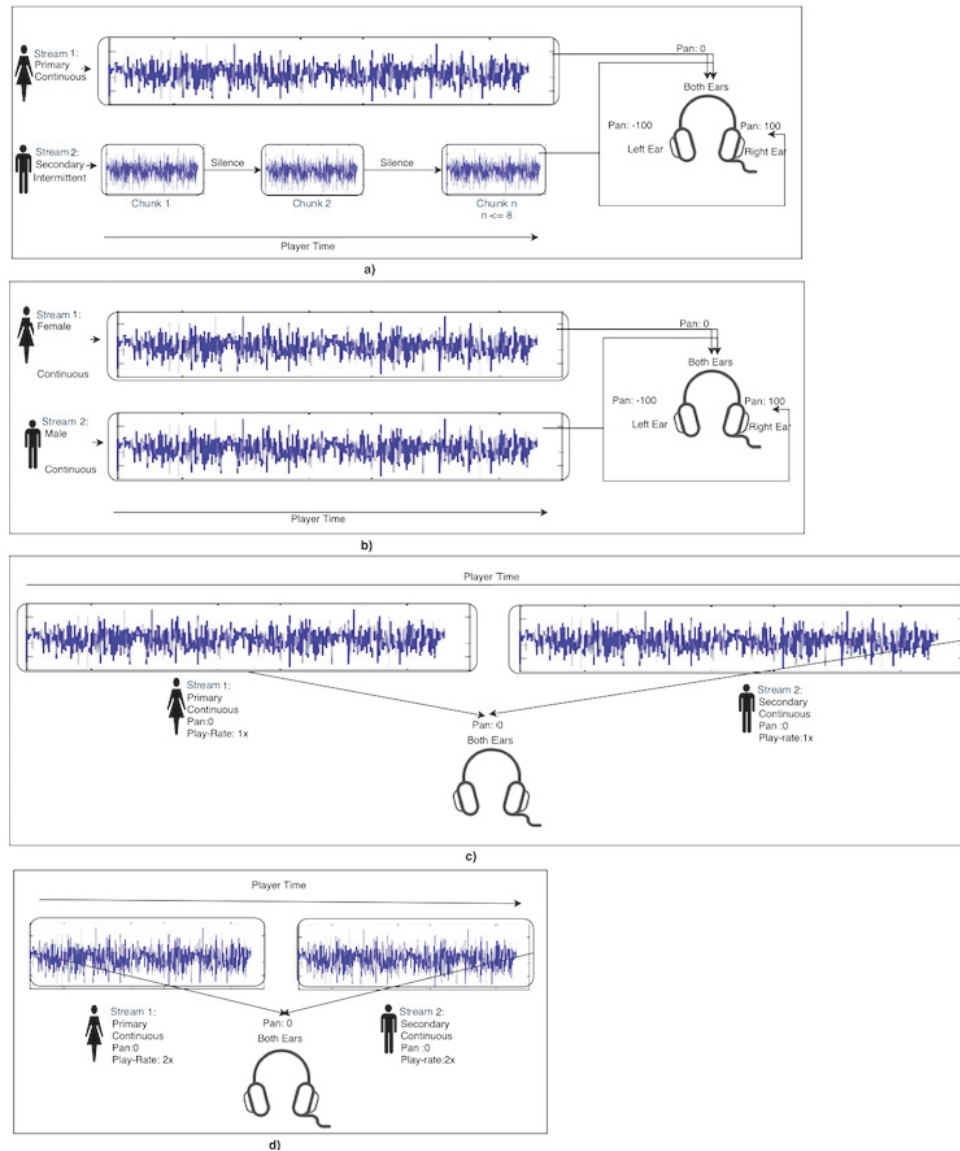


Figure 1: Stimuli Designs: a) Intermittent, b) Continuous, c) Baseline, d) Sequential-2x

3.7 Apparatus

To minimize the participation time for completing the tests and convenience a web-based system using PHP, MySQL, Query, HTML5, CSS, and Bootstrap was designed to play the stimuli. The web system was accessible using latest web browser where 14 HTML audio players each playing one stimulus design were presented on the screen along with the questions under the relevant stimulus player. The tests were conducted in quiet purpose-built creativity and cognition studios (CCS) of the University of Technology, Sydney. Three identical i-Mac computers having 2.7GHz quad-core Intel Core i5 processor, 8 GB RAM, installed with Yosemite 10.10.5 OS were arranged in the studio. To listen to the audio stimuli Beyerdynamic's

DT770 250 OHM headphones were used that were connected to the headphone jack of the computer. Since three computers were used in the studio, therefore, at a time, up to three participants engaged in the experiment simultaneously.

3.8 General Procedure

The selected users were verbally briefed on the study protocol before the start of the experiment, and also the instructions were presented on the screen after registration. Before starting the experiment, users entered their demographic profile information that included, name, age, qualification, first language, country, hearing

impairment and type of computer & headphones used in case of participating from outside of the CCS. At the end of the experiments, user’s subjective response to the concurrent & sequential information communication was also obtained by asking three questions related to user experience. All users’ responses were stored in the MySQL database for the post-experiment analysis.

4 RESULTS

An analysis was carried out on the result data to evaluate speech-based concurrent information designs. For this, users’ comprehension in all concurrent designs were measured and compared with their comprehension in the baseline design (benchmark). For each design, the analysis included two parts: 1) comparing the proportion of users’ responses, and 2) calculating the percentage of correct answers. Afterwards, the intermittent presentation form, identified as the highest-producing comprehension in concurrent designs, was further investigated to assess the underlying mechanism and the users’ comprehension behaviour. The results of Baseline Design Analysis, Concurrent Designs Analysis and the intermittent designs in Detail are individually discussed in following sub-sections.

4.1 Baseline Design Analysis

In the first part of this analysis, the proportion of users response from three options as answers to the questions were determined. It showed how frequently users had selected ‘Don’t Know’ option for both types of ‘Yes’ & ‘No’ expected answers. The analysis showed when the expected answer was ‘No’, 23% responses were selected ‘Don’t know’ by the users which were higher than 8% ‘Don’t Know’ responses in the condition when the expected answer was ‘Yes’. Moreover, the analysis showed an interesting pattern of users inclination towards selecting ‘Yes’ comparing to ‘No’. The percentage, 18%, of selecting ‘Yes’ as a wrong answer was higher than 14% of selecting ‘No’ as a wrong answer. Overall, users selected all three ‘Yes’, ‘No’ and ‘Don’t Know’ options as answers to the questions.

In the second part of the baseline condition analysis, the percentage of the correct answer was calculated to set a benchmark. For this calculation, users response matching to the expected answer counted as a correct answer whereas the opposite answer or the selection of ‘Don’t Know’ option was considered as a wrong answer. The red dot in Figure 3 shows the users correctly answered 65% questions. Inversely, 35% questions either were answered incorrectly, or users didn’t know the answer implying that user could not fully understand the content to answer all the question correctly. Hence, the percentage, (65%) of giving the correct answer in the baseline sequential information communication set the benchmark to compare users’ comprehension in concurrent designs.

4.2 Concurrent Condition Analysis

Following the protocol of performing two types of analysis, the same investigations were performed in the concurrent designs mentioned in Table 1. Since there were two types of contents in concurrent designs, therefore, both types of contents are discussed individually, and showed in Figure 2-(a) & 2-(b). The figure shows that the proportion of selecting ‘Don’t Know’ in DCT concurrent designs remained higher than the baseline conditions. In DCT-Intermittent-Dichotic design, when the expected answer was ‘No’,

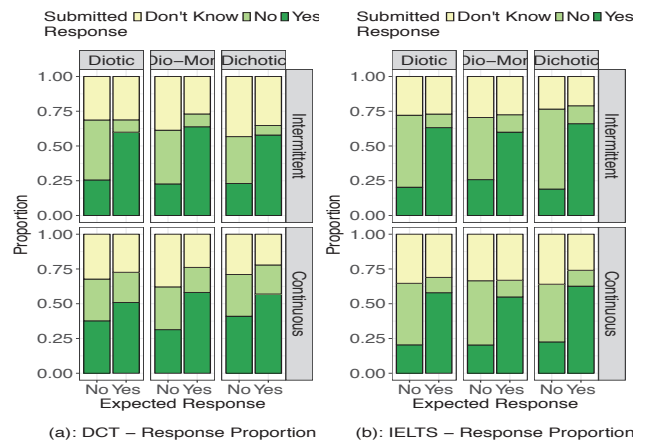


Figure 2: The proportion of users responses.

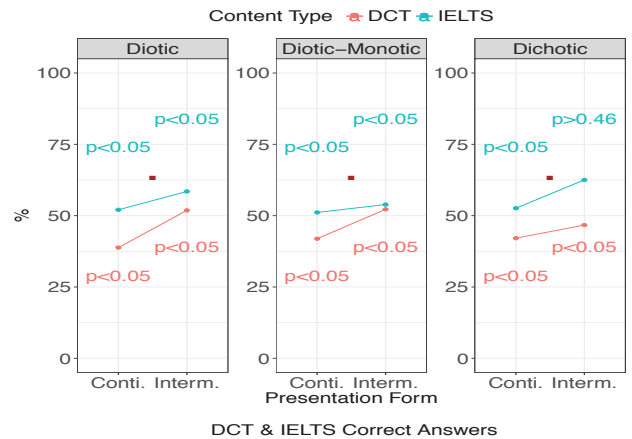


Figure 3: The percentage of correct answers. The p-values against each point show the respective statistical significance against the baseline condition.

the proportion of selecting Don’t know was the highest, i.e. 43% and in the same design when the expected answer was ‘Yes’ the percentage remained 35% which is the second highest proportion. The proportion of selecting “Don’t Know” when the expected answer was ‘No’ remained higher than the percentage when the expected answer was ‘Yes’. Similarly as seen in the baseline condition, the percentage of selecting “Yes”, 23%, as a wrong answer remained higher than the percentage of selecting “No”, 7%, as a wrong answer. The similar trends of proportion have been seen in the other designs as shown in Figure 2-(a). In all the designs based on DCT content, users selected all three ‘Yes’, ‘No’ and ‘Don’t Know’ response as the answers to the questions.

Regarding concurrent designs based on IELTS content, the figure 2-(b) shows the proportion of response submitted by the users. Similar trends appeared in the IELTS as seen in the DCT-based

concurrent designs. However, the proportion of giving correct responses appeared higher and selection of 'don't know' remained lower compared to DCT-based designs. This shows users comprehension remained better in the IELTS-based designs particularly in intermittent designs. In all the designs based on IELTS content, users selected all three 'Yes,' 'No' and 'Don't Know' response as the answers to the questions.

The percentage of correct answers in concurrent designs were also calculated. For comparison, figure 3 shows the percentage of giving correct answers in each concurrent design based on DCT and as well as IELTS contents. In DCT content-based designs, the users' comprehension performance appeared low in comparison to the IELTS content-based designs. The comparison shown in figure 2 reflects that the percentage of giving correct answers was highest in IELTS.Intermittent.Dichotic design, i.e. 63%. This percentage appeared similar ($P = 0.457$) to the baseline benchmark, i.e. 65%. The DCT.Continuous.Diotic design appeared significantly worst design ($P < 0.000$) in communicating concurrent information as the percentage of giving correct answers was as low as 39%. Among both, the continuous and the intermittent forms of concurrent designs, the intermittent form appeared better in communicating speech-based concurrent information.

4.2.1 Intermittent designs in Detail. Since users comprehension performance appeared higher in designs based on intermittent form, therefore, this form was further investigated. In this analysis, the users' behaviour in comprehending 'Competing' questions and its comparison with the comprehension of 'Non-competing' questions in the same design was investigated. For this, those questions in the primary stream of intermittent form-based designs were marked non-competing where the relevant information content to the question was played during the silent interval in the secondary stream. All other questions were marked competing as the content related to those questions was always played in the presence of competing speech. The competing and non-competing marking is visualized in figure 4.

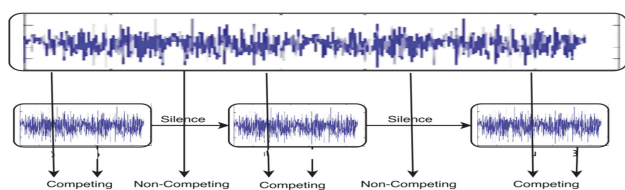


Figure 4: Intermittent Design indicating the 'Competing' and 'Non-competing' questions.

The figures 5 shows the comparison between correct answer percentages in non-competing and competing questions of the DCT & IELTS content-based intermittent designs. The lines and corresponding p-values in figure 5 indicate that the users comprehension was similar in competing and Non-competing' types of questions in DCT content-based intermittent designs. In DCT-Diotic design, for non-competing questions, the percentage remained 56% whereas for competing, the percentage remained 51%. In DCT-Diotic-Monotic the comprehension remained identical. However, in DCT-Dichotic

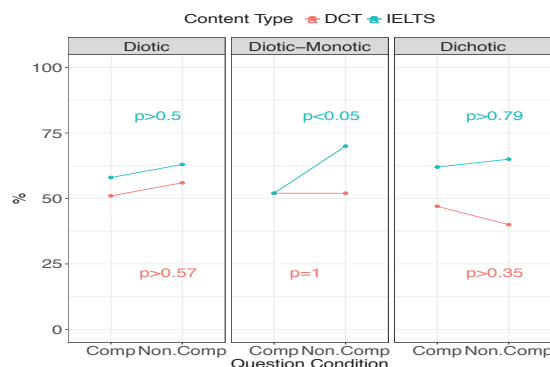


Figure 5: Percentage of giving correct answers w.r.t competing and non-competing questions in intermittent designs. The p-values against each point show the respective statistical significance between the correct answers against question condition.

design, the percentage of the correct answer in non-competing questions remained 40% that was lower than the percentage of correct answers for competing questions i.e. 47%. The graph lines in Figure 5 related to the IELTS content-based intermittent design show that the users comprehension remained slightly higher in terms of correct answers percentage in non-competing questions comparing to the competing questions. In IELTS-Dichotic intermittent design, the correct answer percentage in non-competing question remained 65% whereas in competing questions it remained 62%. The same pattern of slight difference was seen in the other IELTS designs. However, these slight differences in percentage are negligible as the p-Values in figure 5 show that the comprehension difference between both types of questions remained statistically insignificant ($P \geq 0.05$) in all designs except in the case of IELTS.Diotic-Monotic design where p-Value remained $P = 0.018$. Hence, the users' comprehension remained similar in both types of competing and non-competing content presented in the intermittent design.

Furthermore, the intermittent design was a sort of combination of both, the baseline sequential communication and the continuous concurrent communication designs. In the parts of the intermittent design, the portion where the silent intervals in secondary voice appeared, stimulus imitated the sequential communication whereas the other portion where the secondary stream was being played, the speech mocked the continuous communication. In other words, based on the similar information presentation, the non-competing questions shown in the figure 4 were similar to the questions asked in baseline sequential communication, and the competing questions were similar to the questions asked in continuous concurrent communication.

In this analysis, first, the percentage of correct answers in non-competing questions were compared with the questions answered in the baseline condition. Figure 6 (a) shows the percentage for both content types compared with the baseline condition. The analysis showed that the percentage of non-competing questions in IELTS was almost similar, i.e. 66% to the 65% of the baseline condition. However, in DCT the percentage remained significantly

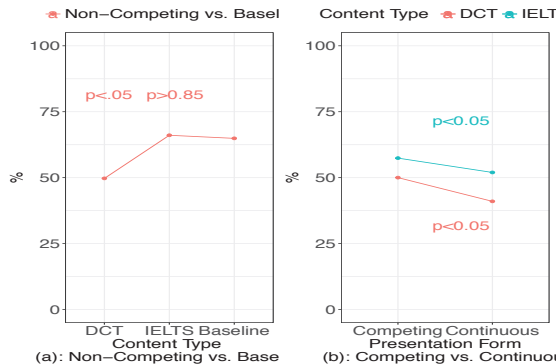


Figure 6: (a) Non-competing questions from concurrent intermittent designs compared with the baseline condition (b) competing questions from concurrent intermittent designs compared with the concurrent continuous designs,

The p-values in a) show the statistical significance of comprehension in presentation forms against the baseline whereas in b) show the statistical significance of comprehension between the competing and continuous question types.

lower ($p < .001$), i.e. 50% as compared to the baseline condition. Since the percentage of correct answers in non-competing was almost similar to the baseline condition in case of IELTS, therefore, this gave an impression that users were not much distracted by the secondary intermittent stream in concurrent condition particularly in IELTS content-based designs.

In the second part, the answers of competing questions in intermittent form-based designs were compared with the answers given in the continuous design of the related content. The figures 6 (b) shows the results. In both the content types, i.e. DCT and IELTS, the correct answer percentage was higher in the competing questions. In DCT comparison, the percentage of giving correct answers for competing questions was 50.31% whereas in the concurrent continuous form the percentage was 40.96%. The similar pattern appeared in IELTS content-based designs where the percentage of the correct answers for the intermittent competing questions remained 57.4% whereas in the concurrent continuous condition it remained 51.96%. The p-value shown in figure 6 indicates the significance difference of giving correct answers between competing and continuous questions. Since $P < 0.05$ in both the types, therefore, comprehension was significantly better in competing questions asked in intermittent concurrent design compared to the continuous concurrent design.

5 DISCUSSION

Figure 3 reflects that the users comprehension was better in intermittent form comparing to the continuous form of delivery. In all six designs, users percentage of correct answers remained higher in case of intermittent comparing to continuous form. It indicates, intermittent based approach provided ample time to the users to understand the context and details of the continuous information that created bandwidth for the user to listen to the intermittent speech

by compromising the continuous speech. This assumption leaves a question, how much information users had obtained and how much they had compromised from both the information streams. An interesting analysis can be carried out to investigate this aspect from the same result dataset.

Besides the intermittent form of presentation, an increased comprehension behaviour was witnessed when the spatial differences were involved additionally in speech-based streams. The designs particularly based on IELTS-content showed that spatial difference played an important role in comprehending the information. The comparison showed in the Figure 2 reflects that the percentage of correct answer was highest. According to the results, the better comprehension comparable to baseline condition can be achieved by providing concurrent information intermittently and in the Dichotic condition.

In figure 2-(a), illustrating the proportion of responses for all designs, the percentage of selecting 'Yes' as a wrong answer remained higher than the percentage of selecting 'No'. Users inclination towards selecting 'Yes' for the higher number of times was based on user's natural instinct towards agreeing with questions when they 'didn't know' the answers. It implies that the absence of 'Don't Know' option could have led to less accurate comprehension calculation.

During analysis, at the results in the designs based on DCT content remained inconsistent compared to the IELTS content-based designs because of the number of factors that include:

- The audio quality of the mono channelled DCT wasn't as much clean in listening as in stereo channelled IELTS.
- The content was natively played in low-pitched (male) voice that was converted into the high-pitched (female) voice for 6 files to attain the discrimination on the basis of gender (fundamental frequency) voice.
- The continuous stories were broken into the chunks to answer the pre-set questions designed natively.

These factors broke the continuity of the discourse/story and audio quality. In case of IELTS these challenges were not faced as a sufficient number of files were available in both the male and the female files, and the audio quality was stereo. Besides this, the questions for IELTS were custom created. Therefore, the broken continuity of the discourse/story challenge didn't appear.

5.1 Limitations & Future Work

This analysis shows the potential of communicating concurrent information with a suitable information design but does not fully cover all the aspects in concurrent condition. This analysis does not adequately cover the user comprehension behaviour. For example, at what point users switched their attention to the secondary stream and how long the attention persisted. During switching the attention, how much information have users lost from the primary information stream while focusing on the secondary stream and vice-versa?

Moreover, as the users had reported about excessive cognitive load, it could be an exciting investigation how many chunks of information could be played as secondary speech-based information in the intermittent design and how long should be the silent intervals between the information chunks. Investigating these limitations

could be an interesting study particularly concerning intermittent design with more details.

REFERENCES

- [1] 1997. Discourse Comprehension Test: Test KIT. http://www.picaprograms.com/discourse_comprehension_test.htm. (1997). [Online; accessed 19-October-2016].
- [2] Jennifer Aydelott, Dinah Baer-Henney, Maciej Trzaskowski, Robert Leech, and Frederic Dick. 2012. Sentence comprehension in competing speech: Dichotic sentence-word priming reveals hemispheric differences in auditory semantic processing. *Language and Cognitive Processes* 27, 7-8 (2012), 1108–1144.
- [3] Jennifer Aydelott, Zahra Jamaluddin, and Stefanie Nixon Pearce. 2015. Semantic processing of unattended speech in dichotic listening. *The Journal of the Acoustical Society of America* 138, 2 (2015), 964–975.
- [4] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller. 2017. Perception of Paralinguistic Traits in Synthesized Voices. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*. ACM, 17.
- [5] Virginia Best, Frederick J Gallun, Antje Ihlefeld, and Barbara G Shinn-Cunningham. 2006. The influence of spatial separation on divided listening a. *The Journal of the Acoustical Society of America* 120, 3 (2006), 1506–1516.
- [6] Karen Church, Mauro Cherubini, and Nuria Oliver. 2014. A Large-scale Study of Daily Information Needs Captured in Situ. *ACM Trans. Comput.-Hum. Interact.* 21, 2, Article 10 (Feb. 2014), 46 pages. DOI: <http://dx.doi.org/10.1145/2552193>
- [7] Andrew RA Conway, Nelson Cowan, and Michael F Bunting. 2001. The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic bulletin & review* 8, 2 (2001), 331–335.
- [8] Ádám Csapó and György Wersényi. 2013. Overview of auditory representations in human-machine interfaces. *ACM Computing Surveys (CSUR)* 46, 2 (2013), 19.
- [9] Alan Dix, Janet E Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-Computer Interaction*. (2003).
- [10] Konstantinos Drossos, Andreas Floros, and Nikolaos-Grigorios Kanellopoulos. 2012. Affective acoustic ecology: Towards emotionally enhanced sound events. In *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*. ACM, 109–116.
- [11] Muhammad Fazal and M Shuaib Karim. 2017. Multiple Information Communication in Voice-Based Interaction. In *Multimedia and Network Information Systems*. Springer, 101–111.
- [12] João Guerreiro. 2013. Using simultaneous audio sources to speed-up blind people's web scanning. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. ACM, 8.
- [13] João Guerreiro. 2016. Towards screen readers with concurrent speech: where to go next? *ACM SIGACCESS Accessibility and Computing* 115 (2016), 12–19.
- [14] João Guerreiro and Daniel Gonçalves. 2014. Text-to-speeches: evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. ACM, 169–176.
- [15] João Guerreiro and Daniel Gonçalves. 2016. Scanning for Digital Content: How Blind and Sighted People Perceive Concurrent Speech. *ACM Transactions on Accessible Computing (TACCESS)* 8, 1 (2016), 2.
- [16] Nandini Iyer, Eric R Thompson, Brian D Simpson, Douglas Brungart, and Van Summers. 2013. Exploring auditory gist: Comprehension of two dichotic, simultaneously presented stories. In *Proceedings of Meetings on Acoustics ICA2013*, Vol. 19. ASA, 050158.
- [17] Philip Kortum. 2008. *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [18] Everdina A Lawson. 1966. Decisions concerning the rejected channel. *The Quarterly journal of experimental psychology* 18, 3 (1966), 260–265.
- [19] Neville Moray. 1959. Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly journal of experimental psychology* 11, 1 (1959), 56–60.
- [20] Cowan Nelson. 1995. Attention and memory: An integrated framework. *Oxford Psychology Series* 26 (1995).
- [21] Jessica A Obermeyer and Lisa A Edmonds. 2018. Attentive Reading With Constrained Summarization Adapted to Address Written Discourse in People With Mild Aphasia. *American journal of speech-language pathology* 27, 1S (2018), 392–405.
- [22] Marie Rivenez, Christopher J Darwin, and Anne Guillaume. 2006. Processing unattended speech. *The Journal of the Acoustical Society of America* 119, 6 (2006), 4027–4040.
- [23] Daisuke Sato, Shaojian Zhu, Masatomu Kobayashi, Hironobu Takagi, and Chieko Asakawa. 2011. Sasayaki: Augmented Voice Web Browsing Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2769–2778. DOI: <http://dx.doi.org/10.1145/1978942.1979353>
- [24] Jonathan H Schuett and Bruce N Walker. 2013. Measuring comprehension in sonification tasks that have multiple data streams. In *Proceedings of the 8th Audio Mostly Conference*. ACM, 11.
- [25] Jonathan H Schuett, Riley J Winton, Jared M Batterman, and Bruce N Walker. 2014. Auditory weather reports: demonstrating listener comprehension of five concurrent variables. In *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound*. ACM, 17.
- [26] Richard J Welland, Rosemary Lubinski, and D Jeffery Higginbotham. 2002. Discourse comprehension test performance of elders with dementia of the Alzheimer type. *Journal of Speech, Language, and Hearing Research* 45, 6 (2002), 1175–1187.

Appendix L

Publication 6 [Submitted]

—, “Investigating Concurrent Speech-based Designs for Efficient Information Communication - Extended Analysis,” *Journal of the Audio Engineering Society (JAES)*, vol. -, no. -, pp. 1–8, 2019, submitted

Investigating Concurrent Speech-based Designs for Efficient Information Communication - Extended Analysis

Muhammad Abu ul Fazal, Sam Ferguson, AND Andrew Johnston,
(Muhammad.AbuUlFazal@uts.edu.au) (Samuel.Ferguson@uts.edu.au) (Andrew.Johnston@uts.edu.au)

University of Technology Sydney, Australia

Speech-based information is usually communicated to users in a sequential manner, but users are capable of obtaining information from multiple voices concurrently. This fact implies that the sequential approach is likely under-utilizing human perception capabilities to some extent and restricting users to perform optimally in an immersive environment. This paper extends the analysis of a comprehensive experiment discussed in [1]. In this paper, we compared the female users and the male users' performance in the concurrent speech-based information communication designs, and also evaluated the comprehension in both, the primary and the secondary, concurrent streams in speech-based concurrent designs. The results showed that both, the female users and the male users, performed similarly, and their comprehension was higher in the primary stream compared to the secondary stream in speech-based concurrent designs.

0 INTRODUCTION

The use of speech in computer interaction seems useful, as humans in their daily life interact with each other using the same method which provides enormous flexibility and efficiency to exchange information. This makes speech an ideal method to be used in auditory displays for communicating information to the user [2]. Conventionally, the auditory displays communicate speech-based information in a single speech stream that under-utilizes human auditory capabilities. Many researchers [3, 4, 5, 6, 7, 8, 2, 9, 10, 11] have worked on introducing concurrent communication through auditory display and show that the humans are capable of noticing, listening and comprehending multiple voice streams simultaneously and that there is potential for communicating multiple information concurrently.

This paper extends speech-based concurrent communication research and investigates the following aims.

1 AIMS & MOTIVATION

1.1 Aims

The aim of this study is to examine designs for speech communication that can communicate concurrent speech-based information similar to the information transfer efficiency that is achieved in conventional sequential speech-based information communication. Additionally, we sought to obtain analyses to satisfy the following ques-

tions: a) Do *Female* and *Male* users perform similarly in comprehending information from the speech-based designs communicating concurrent information? b) How different the comprehension of content remains from the primary and the secondary streams in concurrent communication? c) Which concurrent form presentation, *continuous* or *intermittent*, provides better comprehension? d) Does the spatial difference between the concurrent streams improve concurrent content comprehension?

1.2 Motivation

If this study remains successful, concurrent speech-based communication designs that render better information communication can be adopted in speech-based interaction to communicate more information to listeners in an efficient manner and can help to guide the design of complex and information-heavy speech interaction methods. Concurrent speech can help listen to two TV streams, relevance scanning, scanning for specific information, notifications using a secondary audio channel, TV navigation, and subtitles and assisted navigation, to name a few [12].

2 METHOD

The experiment investigating above aims is outlined below.

2.1 Participants

After receiving institutional Human Research Ethics Committee approval for the research protocol, user participation campaigns were launched. Participants were selected based on two criteria: 1) not having a significant hearing impairment, and 2) having competent English language skills, as the listening experiment's content was in the English language. In total, 34 participants, 14 female and 20 male, took part in the experiment after providing consent. The mean age of the participants was 26 with a standard deviation of 6.

2.2 Design

2.2.1 Concurrent Condition

Within the concurrent condition, initially, six distinct stimuli designs were devised to communicate two speech-based information streams on separate topics concurrently. One stream was in the high pitched (female) voice, and the other was in the low pitched (male) voice. From six, three designs followed the first form, and the remaining three followed the second form of communication from the following list:

- Continuous High-Pitched Stream with Continuous Low-Pitched Stream (Continuous)
- Continuous High-Pitched Stream with intermittent Low-Pitched Stream (Intermittent).

Each of the continuous and intermittent based stimuli design was individually applied with one of the following three pan conditions to involve a spatial difference in streams presenting streams to the specific ear(s):

- 0,0 – Diotic (Both Streams in **both ears**)
- 0,100 – Diotic-Monotic (High-Pitched stream in **both ears** whereas the Low-Pitched stream in the **right ear**)
- -100,100 – Dichotic (High-Pitched stream in the **left ear** whereas the Low-Pitched voice stream in the **right ear**)

All the six design methods were repeated on two types of audio content material that increased concurrent stimuli designs to 12. The audio types of content material were:

- Discourse Comprehension Test (DCT)
- International English Language Testing System (IELTS)

Each of the rendered concurrent stimuli design is described in Table 1 and illustrated in Figure 1 for further clarity.

2.2.2 Baseline Condition

Under this condition, a baseline stimulus representing the conventional speech-based communication was designed where the continuous high pitched information stream followed by a continuous low-pitched information stream was presented sequentially without involving spa-

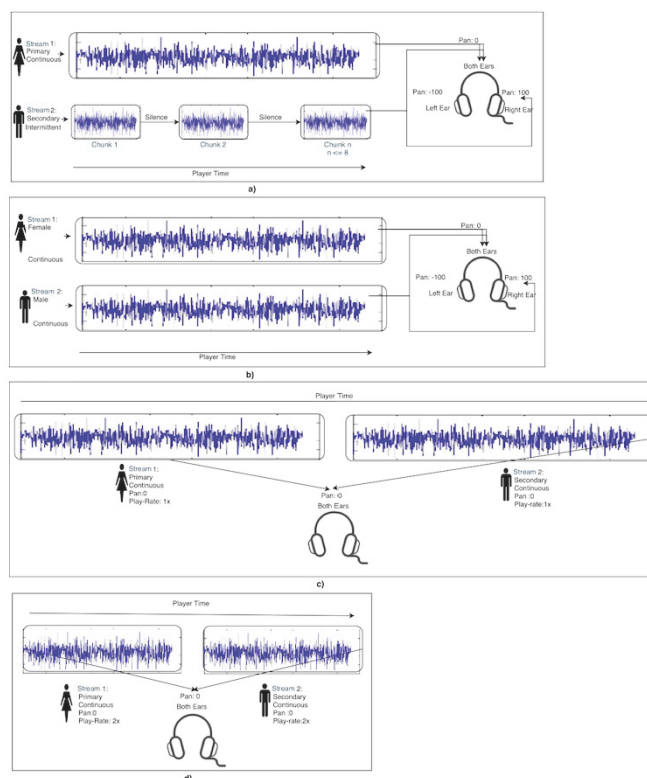


Fig. 1. Stimuli Designs: a) Intermittent, b) Continuous, c) Baseline, d) Sequential-2x

tial difference. The purpose of this design was to determine a benchmark of user comprehension.

Besides baseline stimulus, another sequential stimulus (Seq-2x) was also designed where streams were played following the baseline stimulus method with the only difference in play-rate that was doubled (2x). The purpose of this design was to test another design to communicate multiple information in unit time as shown in Fig. 1-d. This design is not discussed in the result and discussion sections to maintain the simplicity by limiting the scope of this paper.

2.3 Material

For speech-based stimuli designs, two types of content resources were used:

2.3.1 Discourse Comprehension Test (DCT)

The commercially available Discourse Comprehension Test (DCT) [13, 14, 15] is a standardized test to primarily assess the comprehension and retention of spoken narrative discourse by adults suffering from aphasia. The test contains 12 stories where each story having a length from 73 to 95 seconds describes a humorous situation. The material purchased from [16] was received on a CD having twelve mono-channelled audio tracks each presenting a story in the male (low-pitched) voice. To use the stories in the experiment, each track was exported into .wav format with the sample rate of 44.1KHz and the bit rate of 16 using the Apple iTunes software. Since the conceived stimuli designs were to be discriminable by pitch (gender) i.e. high-pitched (female) voice and low-pitched (male) voice, there-

Table 1. Speech-based Concurrent Communication Designs

Concurrent Design			Primary Stream			Secondary Stream		
Content Type	Form	Pan Condition	Voice	Presentation	Ear	Voice	Presentation	Ear
DCT	Continuous	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Both
DCT	Continuous	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Right
DCT	Continuous	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Continuous	Right
DCT	Intermittent	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Both
DCT	Intermittent	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Right
DCT	Intermittent	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Intermittent	Right
IELTS	Continuous	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Both
IELTS	Continuous	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Continuous	Right
IELTS	Continuous	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Continuous	Right
IELTS	Intermittent	Diotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Both
IELTS	Intermittent	Diotic-Monotic	High-Pitched	Continuous	Both	Low-Pitched	Intermittent	Right
IELTS	Intermittent	Dichotic	High-Pitched	Continuous	Left	Low-Pitched	Intermittent	Right

fore, the pitch of the six from twelve tracks was changed by increasing it 17% from the default low frequency male voice using Audacity software (using Sound-touch). This increase in pitch transformed the male (low-pitched) voice into a female (high-pitched) voice. Resultantly, it converted six stories in a high-pitched voice, and six in a low-pitched voice.

2.3.2 International English Language Testing System (IELTS)

The IELTS listening material was also used in the experiment because it was readily available and provided heterogeneous content in stereo-channelled audio files. For the experiment, 12 audio files containing monologue content having the sample rate of 44.1KHz and the bit rate of 16 were selected. In selection, six files were in the male (low-pitched) voice and remaining six were the female (high-pitched) voice. From each monologue file, initial 58-70 seconds of the meaningful content was extracted.

2.4 Stimuli Information

In total, 24 continuous speech-based streams were obtained and processed from both types of material. For having the intermittent streams, the contents of the half of the continuous stream in low-pitched voice were broken into chunks by giving silent intervals of 5 to 10 seconds in them. Each stream was repeatedly applied with each of the three pan conditions 0, 100 and -100 that rendered 72 (24 x 3) streams where 36 were in the high-pitched voice, and 36 (18 continuous and 18 intermittent) were in the low-pitched voice. Then each of the rendered low-pitched stream was repeatedly combined with the high-pitched stream of the same material using the Audacity software for Mac. This multiplication generated 216 combinations to incorporate randomization in the experiment for minimizing the combinational effect in the analysis. From 216 stimuli, randomly 12 (6 DCT + 6 IELTS) were presented to each user where each stimulus was a representation of one of the designs mentioned in table 1. The length of each rendered stimulus was within 55 to 90 seconds except the baseline design. Besides the 12 concurrent designs, the additional two designs, baseline and Seq-2x, were presented to the participants.

2.5 Measures

After listening to each stimulus design, participants answered the questions, discussed in section 2.6, from the stimuli. Since each stimulus was the combination of two streams and each stream had a set of 8 questions, therefore, a user answered 224 questions having yes/no/don't options. The user comprehension was measured on the basis of the number of giving correct answers after listening to each stimulus.

In previous experiments by the authors [10], users often pointed out that they did not know the answer and were looking to select a 'Don't know' option, which wasn't present, and therefore were compelled to choose either 'Yes' or 'No'. This necessarily has resulted in less accurate estimations of user comprehension of the stimulus content, with the assumption being that these participants will naturally choose one of the remaining two options equally. Therefore, in this experimental protocol a third option, 'Don't know' was included, in addition to the usual 'Yes' and 'No' user responses.

2.6 Questionnaire

The DCT material was accompanied with the default questions that were used in the experiment as is, however, for IELTS new questions following the DCT pattern were prepared. Each story had eight questions having yes/no/don't know answers. The questions were arranged in assessment categories to assess the depth of comprehension by the users. For each following category type, two questions were arranged:

- Main Information Stated (MIS)
- Main Information Implied (MII)
- Detailed Information Stated (DTS)
- Detailed Information Implied (DTI)

The questions in MIS were constructed from the main stated information of the story. These questions assessed how much a participant had comprehended the main idea that was repeated or elaborated by other information in the story (main information). The MII questions were based on the information that was not directly discussed in the story, but a user had to infer it from the stated main information.

The questions in DTS were framed from the stated information of the story that estimated the comprehension of detailed information. Detailed information was mentioned only once and not elaborated by other information in the story. DTI questions were based on the information that was not directly explained in the story, but a user had to infer from the detailed information. The implied questions examined whether a user was able to make a mental map or bridging assumptions of the information or not.

2.7 Apparatus

To minimize the participation time for completing the tests and convenience a web-based system using PHP, MySQL, Query, HTML5, CSS, and Bootstrap was designed to play the stimuli. The web system was accessible using latest web browser where 14 HTML audio players each playing one stimulus design were presented on the screen along with the questions under the relevant stimulus player. The tests were conducted in quiet purpose-built creativity and cognition studios (CCS) of the University of Technology, Sydney. Three identical i-Mac computers having 2.7GHz quad-core Intel Core i5 processor, 8 GB RAM, installed with Yosemite 10.10.5 OS were arranged in the studio. To listen to the audio stimuli Beyerdynamic's DT770 250 OHM headphones were used that were connected to the headphone jack of the computer. Since three computers were used in the studio, therefore, at a time, up to three participants engaged in the experiment simultaneously.

2.8 General Procedure

The selected users were verbally briefed on the study protocol before the start of the experiment, and also the instructions were presented on the screen after registration. Before starting the experiment, users entered their demographic profile information that included, name, age, qualification, first language, country, hearing impairment and type of computer & headphones used in case of participating from outside of the CCS. At the end of the experiments, user's subjective response to the concurrent & sequential information communication was also obtained by asking three questions related to user experience. All users' responses were stored in the MySQL database for the post-experiment analysis.

3 RESULTS

The analysis was carried out on the result data to separately evaluate the comprehension by the female users and the male users in speech-based concurrent information designs. For this, comprehension by the female users, and the male users in all concurrent designs were measured and compared with the comprehension in the baseline design (benchmark). For each speech-based concurrent information design, the analysis concerning female and male users included three parts: 1) comparing the proportion of users' responses, 2) calculating the percentage of correct answers,

and 3) comparing the comprehension of content in the primary and the secondary streams.

The results of Baseline Design Analysis, Concurrent Designs Analysis and the Comprehension Comparison between Primary and Secondary Streams are individually discussed in following sub-sections.

3.1 Baseline Design Analysis

In the first part of this analysis, the proportion of responses submitted by the female users and male users from three options as answers to the questions were separately determined in the baseline condition. It showed how frequently users had selected 'Don't Know' option for both types of 'Yes' & 'No' expected answers and also determined the difference between the female and male responses. Regarding female users, the analysis showed when the expected answer was 'No', 26% responses were selected 'Don't know' which were higher than 9% 'Don't Know' responses in the condition when the expected answer was 'Yes'. And in case of male users, 20% responses were selected 'Don't know' when the expected answer was 'No', and 8% responses were selected 'Don't know' when the expected answer was 'Yes'. This shows, both, female and male users, selected all three 'Yes,' 'No' and 'Don't Know' options as answers to the questions.

In the second part of the baseline condition analysis, the percentage of correct answer was calculated to set a benchmark. For this calculation, users response matching to the expected answer counted as a correct answer whereas the opposite answer or the selection of 'Don't Know' option was considered as a wrong answer. Both the female users and the male users answered an equal number of the questions correctly. The red dot in Figure 3, rendering comparison between the baseline and concurrent designs, shows that 65% of the questions were answered correctly by both types of users. Inversely, 35% questions either were answered incorrectly, or users didn't know the answer implying that users could not fully understand the content to answer all the question correctly. Hence, the percentage, (65%) of giving the correct answer in the baseline sequential information communication set the benchmark to compare users' comprehension in concurrent designs.

3.2 Concurrent Condition Analysis

Following the protocol of performing two types of analysis separately for female and male users, the same investigations were performed in the concurrent designs mentioned in Table 1. Since there were two types of contents in concurrent designs, therefore, both types of contents are discussed individually and showed in Figure 2-(a) & 2-(b). The figure shows that the proportion of selecting 'Don't Know' by female and male users in DCT concurrent designs remained higher than the baseline conditions.

The female users in DCT-Intermittent-Dichotic design answered 50% of the questions as 'Don't Know' when the expected answer was 'No', which was the highest proportion of selecting 'Don't Know' in any of the concurrent design. In the same design, when the expected an-

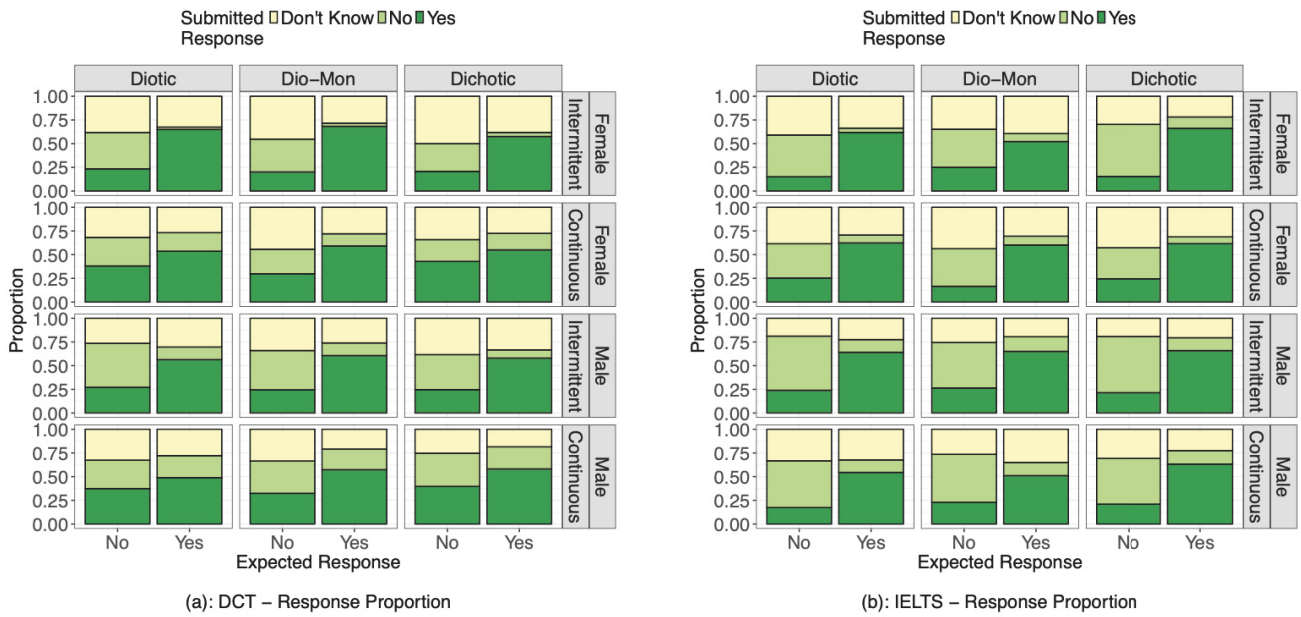


Fig. 2. The Proportion of user responses, grouped by their gender, across all concurrent designs.

answer was 'Yes' the female users answered 38% of questions as 'Don't Know', which was the second highest proportion of selecting 'Don't Know' by the female users in any design. For the female users, the proportion of selecting "Don't Know" when the expected answer was 'No' remained higher than the percentage when the expected answer was 'Yes'. Also, as seen in the baseline condition, the percentage of selecting "Yes", 21%, as a wrong answer remained higher than the percentage of selecting "No", 9%, as a wrong answer by the female users in the same design. As shown in Figure 2-(a), similar trends of proportion by the female users have been seen in all other designs. In all the designs based on DCT content, female users selected all three 'Yes,' 'No' and 'Don't Know' response as answers to the questions.

Similar to the female users, the proportion of answering wrong by male users in DCT-Intermittent-Dichotic design remained the worst, and also, the proportion of selecting "Don't Know" when the expected answer was 'No' remained higher than the percentage when the expected answer was 'Yes'. The male users in DCT-Intermittent-Dichotic design answered 38% of the questions as 'Don't Know' when the expected answer was 'No', and when the expected answer was 'Yes' the male users answered 33% of questions as 'Don't Know', which were the highest proportion of selecting 'Don't Know' by the male users in all designs. Also, the percentage of selecting "Yes", 25%, as a wrong answer remained higher than the percentage of selecting "No", 9%, as a wrong answer by the male users in the same design. As shown in Figure 2-(a), similar trends of proportion by the male users have been seen in all other designs. In all the designs based on DCT content, male users selected all three 'Yes,' 'No' and 'Don't Know' response as the answers to the questions. In all the designs based

on DCT content, similar proportion pattern for answering questions has been seen for both female and male users.

Regarding concurrent designs based on IELTS content, the figure 2-(b) shows the proportion of response submitted by the female and the male users. Similar trends appeared in the IELTS as seen in the DCT-based concurrent designs. However, the proportion of giving correct responses appeared higher and the selection of 'don't know' remained lower compared to DCT-based designs. This shows users comprehension remained better in the IELTS-based designs particularly in intermittent designs. In all the designs based on IELTS content, users selected all three 'Yes,' 'No' and 'Don't Know' response as answers to the questions, and also, similar proportion pattern for answering questions has been seen for both female and male users.

The percentage of correct answers by female and male users in concurrent designs were also calculated separately. For comparison, figure 3 shows the percentage of giving correct answers in each concurrent design based on DCT and as well as IELTS contents by the female and male users. Overall, in DCT content-based designs, the users' comprehension performance appeared low compared to the IELTS content-based designs. And among both, the continuous and the intermittent forms of concurrent designs, the intermittent form appeared better in communicating speech-based concurrent information.

The comparison shown in figure 3 reflects that the percentages of giving correct answers by both, female and male, users were the highest in IELTS.Intermittent.Dichotic design, i.e. 62%, and 63% respectively that shows the percentages of giving correct answers in this design by the female and male users were not different significantly (p value = 0.799). The DCT.Continuous.Diotic design was the worst design in communicating concurrent information to the male users

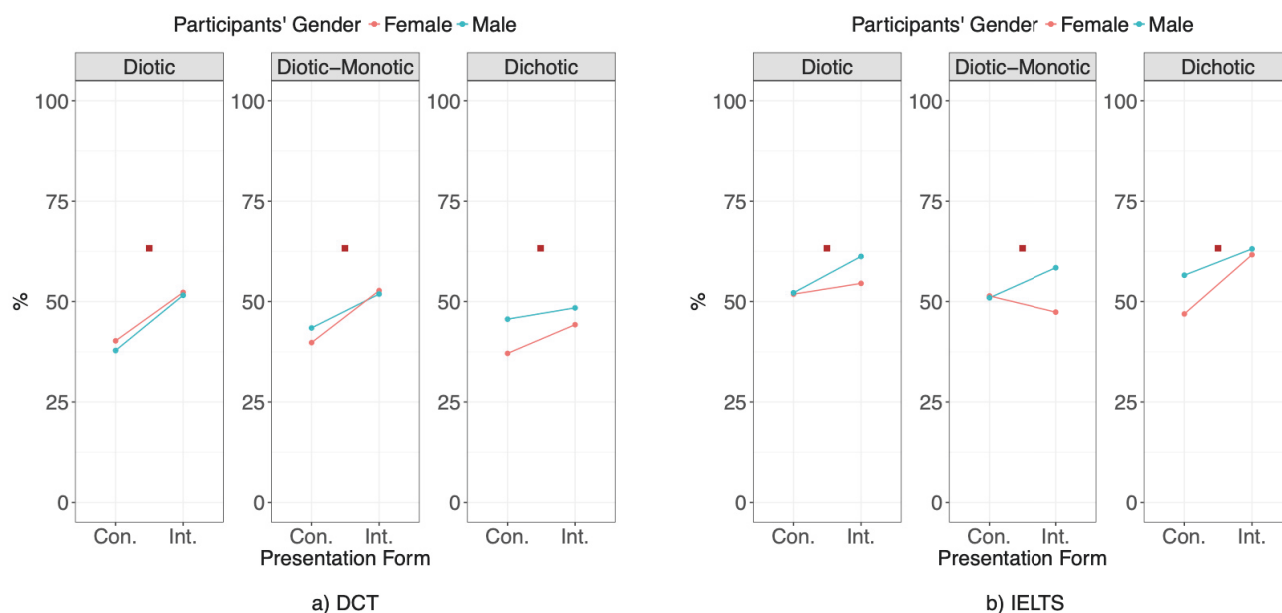


Fig. 3. The percentage of correct answers by users, grouped by their gender, across the concurrent designs.

as the percentage of giving correct answers was as low as 37%. In the same design, the female users answered 40% questions correctly that shows the correct answers given by the female and the male users were not significantly different (p value = 0.632) in this design. For all other designs, the percentages of correct answers by the female and male users were statistically compared using the proportion test, and results in terms of p values are shown in table 2. Results presented in table 2 show that in all the designs, the percentage of giving correct answers by the female users and the male users were not significantly different.

Table 2. Statistical comparison of percentages of correct answers in designs mentioned in Table 1 by the female and male users using the proportion test

Design	p - Value
DCT	
DCT Continuous Diotic	0.6324
DCT Continuous Diotic-Monotic	0.4454
DCT Continuous Dichotic	0.0582
DCT Intermittent Diotic	0.9362
DCT Intermittent Diotic-Monotic	0.9117
DCT Intermittent Dichotic	0.3806
IELTS	
IELTS Continuous Diotic	1.000
IELTS Continuous Diotic-Monotic	0.9855
IELTS Continuous Dichotic	0.0334
IELTS Intermittent Diotic	0.1395
IELTS Intermittent Diotic-Monotic	0.0139
IELTS Intermittent Dichotic	0.7999

3.3 Comprehension Comparison between Streams

In the concurrent speech-based communication designs, two streams were communicated simultaneously to the users, therefore, in this analysis, comprehension compar-

ison between the primary and the secondary streams for each design was conducted for female and the male users separately. The percentage of correct answers by the female and male users in each stream is shown in Figure 4 for each concurrent speech-based design.

In DCT content type, as shown in Figure 4 and p - values mentioned in Table 3, the comprehension by the female and male users remained significantly higher in the primary stream compared to the secondary stream. On account of all the DCT content-based concurrent designs, the average percentages of correct answers in the primary stream remained 56% and 55% whereas in the secondary stream it remained 33% and 37% by the female and male users respectively. In designs based on IELTS content, in three designs - Continuous Diotic, Continuous Dichotic, and Intermittent Dichotic - the comprehension by the female and male users was statistically similar, and in rest of the three designs the comprehension was significantly higher in the primary stream than the secondary stream as shown in Figure 4 and p - values mentioned in Table 3. On account of all the IELTS content-based concurrent designs, the average percentages of correct answers for the primary stream were 57% and 62% whereas in the secondary stream they were 47%, and 52% by the female and male users respectively. In conclusion, for both types of users comprehension mostly remained higher in the primary streams compared to the secondary streams in the speech-based concurrent designs, and also the female users and the male users were able to comprehend the similar amount of information from both concurrent information streams in all speech-based concurrent designs.

4 DISCUSSION

Overall, the female users and the male users performed similarly and were able to comprehend a similar amount of

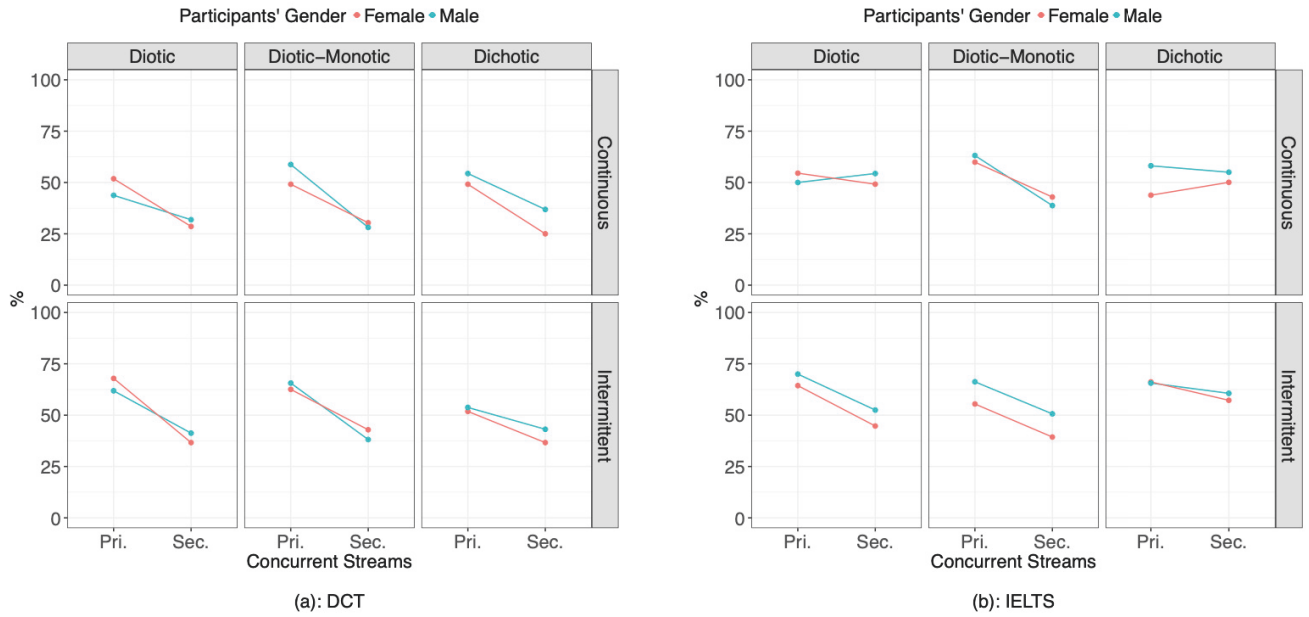


Fig. 4. Percentage of correct answers by users, grouped by their gender, with respect to the primary and the secondary streams in concurrent speech-based designs .

Table 3. Statistical comparison of percentages of correct answers w.r.t to primary and secondary streams in designs mentioned in Table 1 by the female and male users using the proportion test, $p < 0.05 = ***$

Design	Female (p)	Male (p)
DCT		
DCT Continuous Diotic	***	***
DCT Continuous Diotic-Monotic	***	***
DCT Continuous Dichotic	***	***
DCT Intermittent Diotic	***	***
DCT Intermittent Diotic-Monotic	***	***
DCT Intermittent Dichotic	***	0.0735
IELTS		
IELTS Continuous Diotic	0.5036	0.5019
IELTS Continuous Diotic-Monotic	***	***
IELTS Continuous Dichotic	0.4216	0.6519
IELTS Intermittent Diotic	***	***
IELTS Intermittent Diotic-Monotic	***	***
IELTS Intermittent Dichotic	0.216	0.4173

information from both concurrent information streams in all speech-based concurrent designs.

For both types of users, comprehension mostly remained higher in the primary streams compared to the secondary streams in the speech-based concurrent designs. In all intermittent designs and those where information was presented by involving Diotic-Monotic spatial difference, comprehension was better in the primary stream than the secondary stream. In these designs, users considered the female voice (high fundamental frequency) as a primary voice to focus because of the following reasons:

Continuity The primary stream in the female voice was continuous whereas the male voice was played intermittently. The continuity of the stream affected the

user behavior to treat the female voice as a primary voice.

Sound Pressure Level In Diotic-Monotic designs, the female stream was dominant as it was coming to both ears comparing to the male stream that was coming to the right ear only. The difference in sound pressure level (SPL) contributed to treating the female voice as a primary voice to pay attention to.

These findings may help to communicate two streams concurrently, one treated as a primary voice attracting more attention of the users and the other to be treated as secondary information to provide compromisable additional information.

As mentioned in the methods section, in DCT content the default low fundamental frequency (male) voice was increased by 17% to generate the impression that the other stream is being played in a female voice. Considering the DCT.Continuous.Diotic design, where the only difference between both the streams was a difference of values in fundamental frequency, both types of users comprehended more information from the content played in higher fundamental frequency. The result shows that the high-frequency voice attracts more attention of the listeners in case of competing voices compared to the low-frequency voice. The application of this finding could be to use high-frequency voice in a complex sound environment to disseminate the critical information that requires the immediate attention of the listeners among the competing voice-based streams.

For both types of female and male users, comprehension was better in intermittent form comparing to the continuous form of delivery. In all six designs, users percentage of correct answers remained higher in the case of intermittent comparing to the continuous form. It indicates, the intermittent approach provided ample time to the users to under-

stand the context and details of the continuous information that created bandwidth for the users to listen to the intermittent speech by compromising the continuous speech.

Besides the intermittent form of presentation, increased comprehension behavior was witnessed when the spatial differences were involved additionally in speech-based streams. The designs particularly based on IELTS-content showed that spatial difference played an important role in comprehending information. The comparison shown in Figure 3 reflects that the percentage of correct answer was highest by the both, female and male, users. According to the results, the better comprehension comparable to baseline condition can be achieved by both male and female users by providing concurrent information intermittently and in the Dichotic condition.

Both, the female and the male, users selected all three 'Yes,' 'No' and 'Don't Know' options as answers to the questions. In all the concurrent speech-based designs based on DCT and IELTS content, and also in the baseline condition, the similar pattern of proportion for answering questions was seen for both female and male users. However, the proportion of giving correct responses appeared higher and the selection of 'don't know' remained lower compared to DCT-based designs. This shows users comprehension remained better in the IELTS-based designs particularly in intermittent designs.

5 LIMITATIONS & FUTURE WORK

Many users reported a high cognitive load in concurrent speech-based information communication. Since the objective of this study was to assess the content comprehension by the users in concurrent speech-based communication, this experiment required users to listen to content from 14 stimuli designs and answer the questions from the content. The study impacted a high cognitive load and demanded extensive use of memory. In future research, the focus could be on user experience in concurrent communication. An experiment can be designed that could comprehensively investigate the cognitive workload experienced when listening to a variety of combinations of information types and identify the best-suited combinations of information types for concurrent communication.

6 REFERENCES

[1] M. A. u. Fazal, S. Ferguson, A. Johnston, "Investigating Concurrent Speech-based Designs for Information Communication," presented at the *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, AM'18, pp. 4:1–4:8 (2018), [Online]. Available: 10.1145/3243274.3243284.

[2] A. F. Hinde, *Concurrency in auditory displays for connected television*, Ph.D. thesis, University of York (2016).

[3] C. Schmandt, A. Mullins, "AudioStreamer: Exploiting simultaneity for listening," presented at the *Proceedings of the SIGCHI Conference on Human Factors in Com-*

puting Systems, pp. 218–219 (1995), [Online]. Available: 10.1145/223355.223533.

[4] A. T. Mullins, *Audiostreamer: Leveraging The Cocktail Party Effect for Efficient Listening*, Ph.D. thesis, Massachusetts Institute of Technology (1996).

[5] P. Parente, *Clique: Perceptually based, task oriented auditory display for GUI applications*, Ph.D. thesis, The University of North Carolina at Chapel Hill (2008).

[6] J. Guerreiro, D. Goncalves, "Scanning for digital content: How blind and sighted people perceive concurrent speech," *ACM Transactions on Accessible Computing*, vol. 8, no. 1 (2016), [Online]. Available: 10.1109/CVPR.2016.105.

[7] Y. Ikei, H. Yamazaki, K. Hirota, M. Hirose, "vCocktail: multiplexed-voice menu presentation method for wearable computers," presented at the *Virtual Reality Conference*, pp. 183–190 (2006), [Online]. Available: 10.1109/VR.2006.141.

[8] S. Werner, C. Hauck, N. Roome, C. Hoover, D. Choates, "Can VoiceScapes assist in menu navigation?" presented at the *Proceedings of the Human Factors and Ergonomics Society*, vol. 2015, pp. 1095–1099 (2015), [Online]. Available: 10.1177/1541931215591157.

[9] J. A. Towers, *Enabling the Effective Application of Spatial Auditory Displays in Modern Flight Decks*, Ph.D. thesis, The University of Queensland (2016).

[10] M. A. u. Fazal, M. Shuaib Karim, "Multiple information communication in voice-based interaction," in *Advances in Intelligent Systems and Computing*, pp. 101–111 (Springer), [Online]. Available: 10.1007/978-3-319-43982-2_9.

[11] M. A. u. Fazal, S. Ferguson, M. S. Karim, A. Johnston, "Concurrent Voice-Based Multiple Information Communication: A Study Report of Profile-Based Users' Interaction," presented at the *145th Convention of the Audio Engineering Society* (2018).

[12] J. Guerreiro, "Towards screen readers with concurrent speech: where to go next?" *SIGACCESS Accessibility and Computing*, , no. 114, pp. 12–19 (2016).

[13] N. Iyer, E. R. Thompson, B. D. Simpson, D. Brungart, V. Summers, "Exploring auditory gist: Comprehension of two dichotic, simultaneously presented stories," presented at the *Proceedings of Meetings on Acoustics*, vol. 19, pp. 050158–050158 (2013), [Online]. Available: 10.1121/1.4800507.

[14] J. A. Obermeyer, L. A. Edmondsa, "Attentive reading with constrained summarization adapted to address written discourse in people with mild aphasia," *American Journal of Speech-Language Pathology*, vol. 27, no. 1S, pp. 392–405 (2018), [Online]. Available: 10.1044/2017_AJSLP-16-0200.

[15] R. J. Welland, R. Lubinski, D. J. Higginbotham, "Discourse Comprehension Test Performance of Elders With Dementia of the Alzheimer Type," *Journal of Speech Language and Hearing Research*, vol. 45, no. 6, p. 1175 (2002), [Online]. Available: 10.1044/1092-4388(2002/095).

[16]

THE AUTHORS

Appendix M

Publication 7 [Submitted]

M. A. u. Fazal, S. Ferguson, and A. Johnston, "Evaluation of Information Comprehension in Speech-based Designs for Concurrent Audio Streams," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. -, no. -, pp. 1–18, 2018, submitted

Evaluation of Information Comprehension in Speech-based Designs for Concurrent Audio Streams

MUHAMMAD ABU UL FAZAL, Creativity & Cognition Studios, School of Software, Faculty of Engineering & IT, University of Technology, Sydney, Australia

SAM FERGUSON, Creativity & Cognition Studios, School of Software, Faculty of Engineering & IT, University of Technology, Sydney, Australia

ANDREW JOHNSTON, Creativity & Cognition Studios, School of Software, Faculty of Engineering & IT, University of Technology, Sydney, Australia

In human-computer interaction, particularly in multimedia delivery, information is communicated to users sequentially, but users are capable of receiving information from multiple sources concurrently. This mismatch indicates that a sequential mode of communication does not utilize human perception capabilities as efficiently as possible. This paper reports an experiment that investigated various speech-based (audio) concurrent designs and evaluated the comprehension depth of information by comparing comprehension performance across several different formats of question (main/detailed, implied/stated). The results showed that users, besides answering the main questions, were also successful in answering the implied questions, as well as the questions that required detailed information, and that the pattern of comprehension depth remained similar to that seen to a baseline condition, where only one speech source was presented. However, the participants answered more questions correctly that were drawn from the main information, and performance remained low where the questions were drawn from detailed information. The results are encouraging to explore the concurrent methods further for communicating the multiple information streams efficiently in human-computer interaction including the multimedia.

CCS Concepts: • **Information systems** → **Multimedia streaming**; • **Human-centered computing** → *Empirical studies in HCI*;

Additional Key Words and Phrases: Concurrent Audio, Audio Streams, Auditory displays, Streaming, Voice-based Interaction, Concurrent Speech-based Information Comprehension, Audio Spatial Location, Intermittent & Continuous Audio Presentation, Comprehension Depth

ACM Reference Format:

Muhammad Abu ul Fazal, Sam Ferguson, and Andrew Johnston. 0. Evaluation of Information Comprehension in Speech-based Designs for Concurrent Audio Streams. *ACM Trans. Multimedia Comput. Commun. Appl.* 0, 0, Article 0 (0), 18 pages. <https://doi.org/0000001.0000001>

1 INTRODUCTION

In this era of increasing ubiquitous computing, many people rely on computer systems including mobile telephony to complete their daily tasks [32, 33]. Usually, they use a visual interface to interact with a system, but given the high volume of information that needs to be consumed, users

Authors' addresses: Muhammad Abu ul Fazal, Creativity & Cognition Studios, School of Software, Faculty of Engineering & IT, University of Technology, Sydney, Broadway, Sydney, NSW, 2007, Australia, Muhammad.AbuUlFazal@uts.edu.au; Sam Ferguson, Creativity & Cognition Studios, School of Software, Faculty of Engineering & IT, University of Technology, Sydney, NSW, Australia, Samuel.Ferguson@uts.edu.au; Andrew Johnston, Creativity & Cognition Studios, School of Software, Faculty of Engineering & IT, University of Technology, Sydney, NSW, Australia, Andrew.Johnston@uts.edu.au.

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

© 0 Association for Computing Machinery.

1551-6857/0/0-ART0 \$15.00

<https://doi.org/0000001.0000001>

sometimes find that a visual interface is insufficient to service this need. Fortunately, there are other modalities that facilitate human-computer interaction either by supporting or as an alternative to the visual interface. Voice-based interfaces are emerging as an important one.

Voice-based interaction facilitates users to interact with a system in hands-free, eyes-free contexts [4, 5, 30]. The auditory modality provides a useful method that uses the aural channel of the user to communicate information. Due to its high sensitivity [49] and omnidirectional nature, the auditory channel facilitates users to listen or get a gist of the information from multiple sources within their surroundings concurrently [12]. In the real world, this so-called "cocktail party effect" is achieved by having the advantage of the binaural difference cues [8, 11].

Since users can listen and comprehend multiple streams of information concurrently through auditory perception and are also challenged with expanded information needs [13], one plausible solution to maximize information communication is to provide multiple streams of speech-based information concurrently in natural voices or realistic voice synthesizers [36]. But most of the present voice-based interaction designs provide communication in a sequential form to the user, which theoretically underutilizes natural human perception capabilities [15, 16], as the single channel acts as a bottleneck for the amount of information that can be communicated in a given time [43].

This study determines the viability of communicating concurrent information streams by examining different concurrent speech-based designs that can optimally exploit the human auditory capabilities and deliver information efficiently. This paper is arranged as follows. The background followed by aims & motivation and a method of the experiment are presented. An analysis of the results of experiment and discussion are followed by a short conclusion.

2 BACKGROUND

These days, the internet has become an essential tool to seek information [50], where huge growth of multimedia data is being witnessed [48]. Such multimedia data include broadcast news, movies, TV programs, lectures, music and many other types of data, which is playable in the form of audio or video streams [7]. The digital streaming enables a user to watch or listen to the content using any sized computer and bandwidth [27, 41]. Some of the popular streaming based services are YouTube, Twitch, Netflix, Spotify, SoundCloud etc. These services provide flexibility and personal freedom to the users to access information or entertainment according to their needs [17]. Since the inception of the digital audio and video, a lot of efforts and developments have been made to improve the efficiency, scalability, and the adaptability of the streams [20]. After the penetration of the streaming giants like YouTube, Netflix, Spotify, SoundCloud, etc., streaming is being considered as an industry where the challenge is, how the information could be delivered to the users with the acceptable quality and pace that user is looking for [26, 28].

These days, users have huge information needs that they want to seek to meet the various challenges of the life, and also to remain updated with the events happening in their surroundings. To fulfill these needs, users listen to digital streams to either get a deep understanding or to get the quick gist related to a particular matter. For example, a student having a short time or patience would prefer to skip various parts of the content to reach to the specific part of the information or would listen to the content at high playback-rate [40]. Therefore, the high-playback speed option to listen to the information quickly is getting popular. The popular platforms like YouTube, Udacity, edX provide users with the options to set the playback-rate according to their need [40].

The high playback-rate is one of the methods to communicate the information quickly. The other design could be to play two information streams on different topic concurrently to the users with the regular playback-rate, as the users are capable of listening and comprehending multiple information concurrently through auditory perception [13]. Recent investigations on

speech-based voices have reported promising performance by participants in listening to two streams presented concurrently. Psychological studies [14, 31, 37, 38, 42] have shown that humans can listen and process information that exists outside of the immediate auditory focus. A listener can selectively read out ‘secondary’ information from working memory [6, 14]. This selective readout from concurrent information sources is aided by various audio signal cues along with the users’ personal and contextual circumstances [2, 3].

Many studies have been published on the perceptual and cognitive mechanisms for auditory processing and stream segregation, but the underlying mechanism is yet not fully known [18, 21, 34]. Many psychologists and philosophers believe that the mental representation of the surroundings develops from the processing of information provided by the human senses [9]. Auditory Scene Analysis (ASA), coined by the renowned psychologist Albert Bregman, is considered a foundational model to understand auditory perception [10]. ASA explains how the auditory system detects and separates the multiple complex waveforms into the meaningful representations [35]. ASA, based on gestalt laws of grouping [47], states that the similarity and difference in cues such as proximity, temporal Proximity, similarity, continuation, familiarity and belongingness help in perceiving the sound in streams and also separating one stream from another.

In voice-based human-computer interaction, [Fazal and Karim](#) examined a design method to simultaneously communicate two speech-based information streams through an experimental study [19]. The results showed that concurrent information communication is possible using voice in human-machine interaction. Users were not only able to discriminate between the two information streams, but they were also able to get concurrent multiple information meaningfully in a shorter time using their selection and attention abilities. The study reported that users showed interest in concurrent multiple information communication. [Schmandt and Mullins](#) introduced AudioStreamer tool that exploits peoples ability to separate the two streams into distinct sources for effective browsing from the multiple concurrent streams of real-time or stored audio. [Vazquez Alvarez and Brewster](#) tested a continuous podcast competed with an audio menu concurrently using divided-attention abilities of users and showed that the spatial audio increases the users’ ability to attend two streams concurrently. The results also showed that the dividing attention had a significant effect on overall performance.

Moreover, [Iyer et al. \(2013\)](#) carried out experiments to understand the amount of the information comprehension, and also the nature of the semantic processing in concurrent information communication. The results of these experiments identified that participants were able to apprehend the main idea of the *unattended* story to a level that was higher than chance. The outcome of these experiments appeared consistent with studies of visual gist processing that suggest that the auditory system receives global features before diverting full attention to the stream. Regarding scanning information from the concurrent sources, [Guerreiro and Gonçalves](#) carried out experiments with sighted and with visually impaired persons to determine people’s ability to find important speech content from two, three, or four speech channels played concurrently [22, 24, 25]. The study established: (1) Both the sighted and visually impaired users could successfully scan information from concurrent speech-based streams with no significant difference in their performance. (2) Two concurrent voices render better results in scanning and selecting relevant information than three, and even further, four. (3) The use of both two and three simultaneous sources depends on the task intelligibility demands and listener capabilities (working memory). (4) Gender difference in voices doesn’t play a role in the higher understanding of the content; however, it is highly demanded by the users. (5) The spatial difference in sources appeared the best cue in concurrent speech [11].

3 AIMS & MOTIVATION

3.1 Aims

This paper aims to evaluate comprehension depth of both the primary and the secondary information streams in speech-based concurrent information designs and determine which design remains the most effective in communicating speech-based information concurrently.

3.2 Motivation

The motivation is to come up with speech-based concurrent information design for auditory displays that could efficiently communicate concurrent information nearly equal to the performance that people achieve in conventional sequential information communication. This study can also help to guide the design of complex and information-heavy speech interaction methods. The quick communication of multiple information streams can be helpful in listening to digital streams, relevance scanning, scanning for specific information, notifications using a secondary audio channel, navigations, etc. [23].

To achieve this goal, we designed an experiment that played two information streams concurrently with different design configurations.

4 METHOD

The standardized experiment discussed in this paper is an extension of the work carried out by [Fazal and Karim](#), [Iyer et al.](#), and [Guerreiro and Gonçalves](#). The method adopted for this experiment is outlined below.

4.1 Participants

After receiving institutional Human Research Ethics Committee approval for the research protocol, user participation campaigns were launched. The participants were selected based on two criteria: 1) not having a significant hearing impairment, and 2) having competent English language skills, as the listening experiment's content was in the English language. In total, 34 participants, 14 female, and 20 male took part in the experiment after providing consent. The mean age of the participants was 26 with the standard deviation of 6.

4.2 Design

4.2.1 Concurrent Condition.

Independent Variables. We manipulated three independent variables within the designs:

Content Type: The stimuli were either created from mono-channeled DCT or the stereo channeled IELTS audio content files.

Presentation Form: The presentation form was either concurrent with two *continuous streams*, or concurrent with one continuous stream and another intermittent.

Spatial Configuration: The spatial configuration that the stimuli were presented in was one of three options: Diotic, Diotic-Monotic or Dichotic.

Within the concurrent condition, initially, six distinct stimuli designs were devised to communicate two speech-based information streams on separate topics concurrently. One stream was in the female (high fundamental frequency) voice, and the other was in the male (low fundamental frequency) voice. From six, three designs followed the first form, and the remaining three followed the second form of communication from the following list:

- Continuous Female Stream with Continuous Male Stream (Continuous)
- Continuous Female Stream with Intermittent Male Stream (Intermittent)

Table 1. Speech-based Concurrent Communication Designs

Concurrent Design			Primary Stream			Secondary Stream		
Content	Form	Pan	Voice	Presentation	Ear	Voice	Presentation	Ear
DCT								
DCT	Continuous	Diotic	Female	Continuous	Both	Male	Continuous	Both
DCT	Continuous	Dio-Mon	Female	Continuous	Both	Male	Continuous	Right
DCT	Continuous	Dichotic	Female	Continuous	Left	Male	Continuous	Right
DCT	Intermittent	Diotic	Female	Continuous	Both	Male	Intermittent	Both
DCT	Intermittent	Dio-Mon	Female	Continuous	Both	Male	Intermittent	Right
DCT	Intermittent	Dichotic	Female	Continuous	Left	Male	Intermittent	Right
IELTS								
IELTS	Continuous	Diotic	Female	Continuous	Both	Male	Continuous	Both
IELTS	Continuous	Dio-Mon	Female	Continuous	Both	Male	Continuous	Right
IELTS	Continuous	Dichotic	Female	Continuous	Left	Male	Continuous	Right
IELTS	Intermittent	Diotic	Female	Continuous	Both	Male	Intermittent	Both
IELTS	Intermittent	Dio-Mon	Female	Continuous	Both	Male	Intermittent	Right
IELTS	Intermittent	Dichotic	Female	Continuous	Left	Male	Intermittent	Right

Each of the Continuous and Intermittent based stimuli design was individually applied with one of the following three pan conditions to involve a spatial difference in streams presenting streams to the specific ear(s):

- 0,0 – Diotic (Both Streams in **both ears**)
- 0,100 – Diotic-Monotic (Female stream in **both ears** whereas the Male stream in the **right ear**)
- -100,100 – Dichotic (Female stream in the **left ear** whereas the Male voice stream in the **right ear**)

All the six design methods were repeated on two types of audio content material that increased concurrent stimuli designs to 12. The audio types of content material were:

- Discourse Comprehension Test (DCT)
- International English Language Testing System (IELTS)

Each of the rendered concurrent stimuli design is described in Table 1 and illustrated in Figure 1 for further clarity.

4.2.2 Baseline Condition. Under this condition, a baseline stimulus representing the conventional speech-based communication was designed where the Continuous female information stream followed by a Continuous male information stream was presented sequentially without involving spatial difference. The purpose of this design was to determine a benchmark of user comprehension in the baseline condition that could subsequently be used to evaluate users' comprehension in concurrent condition.

Besides baseline stimulus, another sequential stimulus (Seq-2x) was also designed where streams were played following the baseline stimulus method with the only difference in play-rate that was doubled (2x). The purpose of this design was to test another design to communicate multiple information in unit time as shown in Fig. 1-d. This design is not discussed in the result and discussion sections to maintain the simplicity by limiting the scope of this paper.

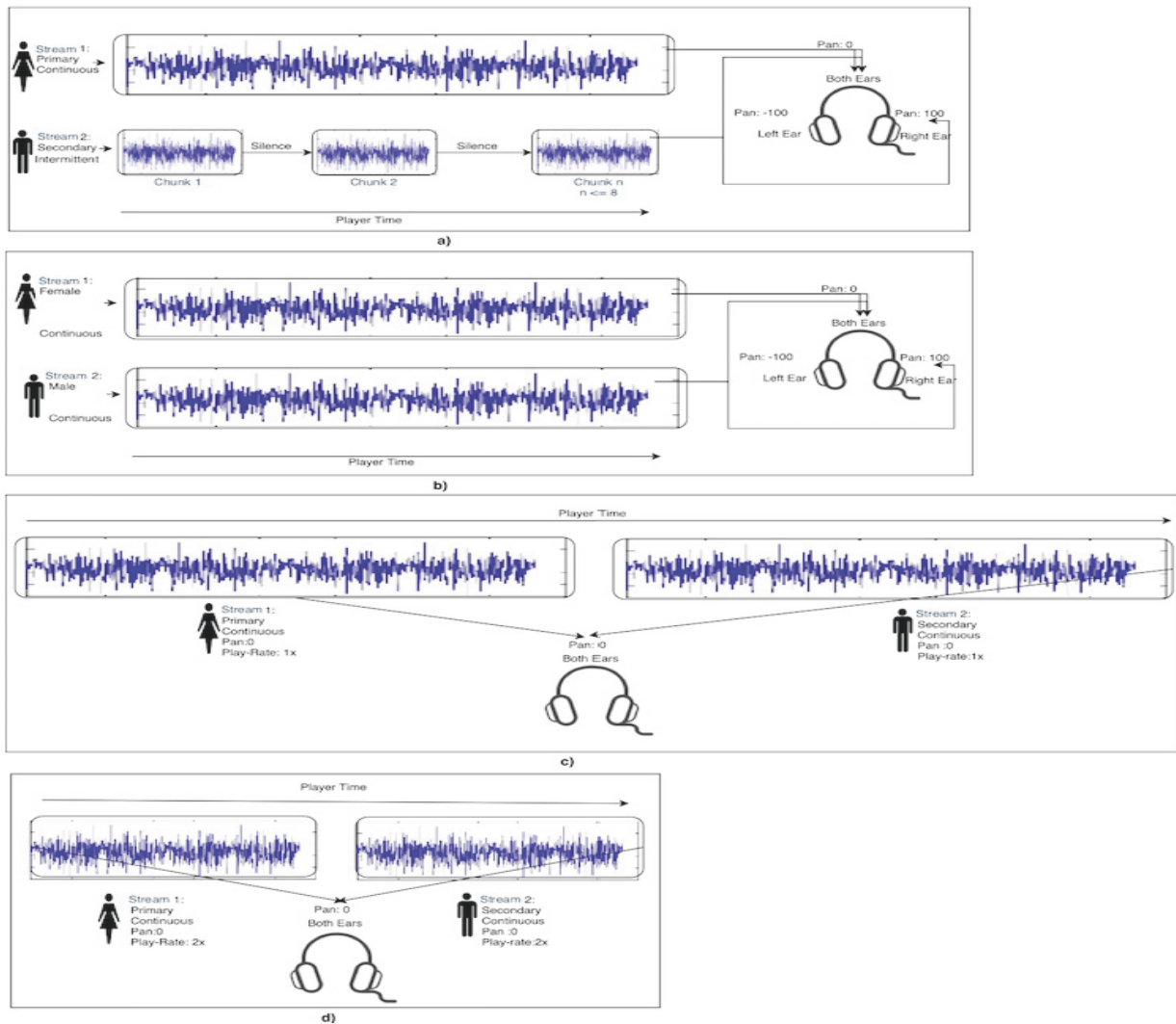


Fig. 1. Stimuli Designs: a) Intermittent: Continuous Stream in Female Voice with Intermittent Stream in Male Voice having Panning / Spatial Conditions of 0-0 (Diotic), 0-100 (Diotic-Monotic), (-)100-100 (Dichotic), b) Continuous: Continuous Stream in Female Voice with Continuous Stream in Male Voice having Panning / Spatial Conditions of 0-0 (Diotic), 0-100 (Diotic-Monotic), (-)100-100 (Dichotic), c) Baseline: Continuous Stream in Female Voice followed by Continuous Stream in Male Voice with no Panning Conditions (0), d) Sequential-2x Stimulus: Continuous Stream in Female Voice followed by Continuous Stream in Male Voice with doubled play-rate

4.3 Material

For speech-based stimuli designs, two types of content resources were used:

4.3.1 Discourse Comprehension Test (DCT). The commercially available Discourse Comprehension Test (DCT) [29, 39, 46] is a standardized test to primarily assess the comprehension and retention of spoken narrative discourse by adults suffering from aphasia. The test contains 12 stories where each story having a length from 73 to 95 seconds describes a humorous situation. The material purchased from [1] was received on a CD having twelve mono-channelled audio tracks each presenting a story in the male (low fundamental frequency) voice. To use the stories in the experiment, each track was exported into .wav format with the sample rate of 44.1KHz and the bit rate of 16 using the Apple iTunes software. Since the conceived stimuli designs were to

be discriminable by gender (fundamental frequency) i.e. female (high frequency) voice and male (low frequency) voice, therefore, the pitch of the six from twelve tracks was changed by increasing it 17% from the default low frequency male voice using Audacity software (using Sound-touch). This increase in pitch transformed the male voice into a female voice. Resultantly, it converted six stories in a female voice, and six in a male voice.

4.3.2 International English Language Testing System (IELTS). The IELTS listening material was also used in the experiment because it was readily available and provided heterogeneous content in stereo-channelled audio files. For the experiment, 12 audio files containing monologue content having the sample rate of 44.1KHz and the bit rate of 16 were selected. In selection, six files were in the male (low fundamental frequency) voice and remaining six were the female (low fundamental frequency) voice. From each monologue file, initial 58-70 seconds of the meaningful content was extracted.

4.4 Stimuli Information

In total, 24 Continuous speech-based streams were obtained and processed from both types of material. For having the Intermittent streams, the contents of the half of the Continuous stream in male voice were broken into chunks by giving silent intervals of 5 to 10 seconds in them. Each stream was repeatedly applied with each of the three pan conditions 0, 100 and -100 that rendered 72 (24 x 3) streams where 36 were in the female voice, and 36 (18 Continuous and 18 Intermittent) were in the male voice. Then each of the rendered male stream was repeatedly combined with the female stream of the same material using the Audacity software for Mac. This multiplication generated 216 combinations to incorporate randomization in the experiment for minimizing the combinational effect in the analysis. From 216 stimuli, randomly 12 (6 DCT + 6 IELTS) were presented to each user where each stimulus was a representation of one of the designs mentioned in table 1. The length of each rendered stimulus was within 55 to 90 seconds except the baseline design. Besides the 12 concurrent designs, the additional two designs, baseline and Seq-2x, were presented to the participants.

4.5 Measures

After listening to each stimulus design, participants answered the questions, discussed in section 4.6, from the stimuli. Since each stimulus was the combination of two streams and each stream had a set of 8 questions, therefore, a user answered 224 questions having yes/no/don't options. The user comprehension was measured on the basis of the number of the correct answers after listening to each stimulus.

In previous experiments by the authors [19], users often pointed out that they did not know the answer and were looking to select a 'Don't know' option, which wasn't present, and therefore were compelled to choose either 'Yes' or 'No'. This necessarily has resulted in less accurate estimations of user comprehension of the stimulus content, with the assumption being that these participants will naturally choose one of the remaining two options equally. Therefore, in this experimental protocol a third option, 'Don't know' was included, in addition to the usual 'Yes' and 'No' user responses.

4.6 Questionnaire

The DCT material was accompanied with the default questions that were used in the experiment as is, however, for IELTS new questions following the DCT pattern were prepared. Each story had eight questions having yes/no/don't know answers. The questions were arranged in assessment

categories to assess the depth of comprehension by the users. For each following category type, two questions were arranged:

- Main Information Stated (MIS)
- Main Information Implied (MII)
- Detailed Information Stated (DTS)
- Detailed Information Implied (DTI)

The questions in MIS were constructed from the main stated information of the story. These questions assessed how much a participant had comprehended the main idea that was repeated or elaborated by other information in the story (main information). The MII questions were based on the information that was not directly discussed in the story, but a user had to infer it from the stated main information. The questions in DTS were framed from the stated information of the story that estimated the comprehension of detailed information. Detailed information was mentioned only once and not elaborated by other information in the story. DTI questions were based on the information that was not directly explained in the story, but a user had to infer from the detailed information. The implied questions examined whether a user was able to make a mental map or bridging assumptions of the information or not. A couple of IELTS streams' content and associated questions are mentioned in Appendix-A.

4.7 Apparatus

To minimize the participation time for completing the tests and convenience a web-based system using PHP, MySQL, JQuery, HTML5, CSS, and Bootstrap was designed to play the stimuli. The web system was accessible using latest web browser where 14 HTML audio players each playing one stimulus design were presented on the screen along with the questions under the relevant stimulus player. The tests were conducted in quiet purpose-built creativity and cognition studios (CCS) of the University of Technology, Sydney. Three identical i-Mac computers having 2.7GHz quad-core Intel Core i5 processor, 8 GB RAM, installed with Yosemite 10.10.5 OS were arranged in the studio. To listen to the audio stimuli Beyerdynamic's DT770 250 OHM headphones were used that were connected to the headphone jack of the computer. Since three computers were used in the studio, therefore, at a time, up to three participants engaged in the experiment simultaneously.

4.8 General Procedure

The selected users were verbally briefed on the study protocol before the start of the experiment, and also the instructions were presented on a screen after registration. Before starting the experiment, users entered their demographic profile information that included, name, age, qualification, first language, country, hearing impairment and type of computer & headphones used in case of participating from outside of the CCS. At the end of the experiments, user's subjective response to the concurrent & sequential information communication was also obtained by asking three questions related to user experience. All users' responses were stored in the MySQL database for the post-experiment analysis.

5 RESULTS

An analysis was carried out that started by comparing the overall comprehension of content in the primary stream with the overall comprehension of content in secondary streams and then extended to determine the depth of comprehension in each stream in speech-based concurrent information designs. The comprehension depth is defined as understanding spoken information from the scale of main to the detailed levels. The arrangement of questions in MIS, MII, DTS, and DTI categories and users' correct answers to these categorised questions determined the comprehension depth in both

the primary and the secondary streams individually in each stimulus design. The analysis regarding comprehension depth started with the baseline condition and estimated users comprehension to set a benchmark which was subsequently used to compare users' comprehension depth during the analysis of concurrent designs. The analysis also covered the qualitative response submitted by the users explaining experience particularly emphasizing the cognitive load incurred when interacting with the speech-based concurrent information designs.

The results of overall comprehension comparison between primary and secondary streams for each design followed by the comprehension depth results for Baseline Design and the Concurrent Designs are discussed in sub-sections 5.1, 5.2 & 5.3. The Qualitative Response Submitted by the Users is discussed in sub-section 5.4.

5.1 Overall Comprehension Comparison between Streams

In the first part of this analysis, overall comprehension comparison between the primary and the secondary streams for each design was conducted. To reflect the comparison, the percentage of correct answers in each stream is shown in Figure 2 for each concurrent speech-based design. The users' response to a question that matched to the expected answer counted as a correct answer whereas the opposite answer or the selection of 'Don't Know' option was considered as a wrong answer. In DCT content type, as shown in Figure 2 and p-values mentioned in Table 2, the comprehension remained significantly higher in the primary stream in comparison to the secondary stream. The comprehension difference remained unchanged in both the Continuous and the Intermittent forms of the concurrent speech-based designs. On account of all the DCT content-based concurrent designs the average percentage for the primary stream remained 56% whereas in the secondary stream it remained 35%. In designs based on IELTS content, in three designs - Diotic Continuous, Dichotic Continuous and Dichotic Intermittent - the comprehension remained the same, and in rest of the three designs the comprehension remained higher in the primary stream than the secondary stream as shown in Figure 2 and p-values mentioned in Table 2. On account of all the IELTS content-based concurrent designs the average percentage for the primary stream remained 60% whereas in the secondary stream it remained 50%. In conclusion, users comprehension mostly remained higher in the primary streams compared to the secondary streams in the speech-based concurrent designs.

5.2 Comprehension Depth for Baseline Condition

The second part of the analysis started with an evaluation of the comprehension depth of the content in baseline condition. The individual percentage with respect to MIS, MII, DTS, and DTI was calculated to set a benchmark that was later used to compare the comprehension in the concurrent design. Figure 3 using red line shows the analysis of the Baseline condition. The percentage of correct answers of the questions set from MIS remained 85% whereas in MII, DTS, DTI, it remained 72%, 51%, 51% respectively. The analysis shows that the comprehension in MIS remained significantly higher comparing to the other information types.

5.3 Comprehension Depth for Concurrent Condition

Following the pattern adopted in analysis for baseline condition the concurrent depth was evaluated for each concurrent speech-based design. Figure 3 shows the percentage of correct answers in both the streams w.r.t MIS, MII, DTS, DTI for each design individually. The each MIS, MII, DTS, DTI data point (percentage) of each concurrent speech-based design is statistically compared with relevant data point in the benchmark set from the baseline condition and mentioned in Table 3. Almost in all the speech-based concurrent designs the comprehension remained significantly lower than the benchmark except in one design i.e. IELTS.Intermittent.Dichotic. In the primary

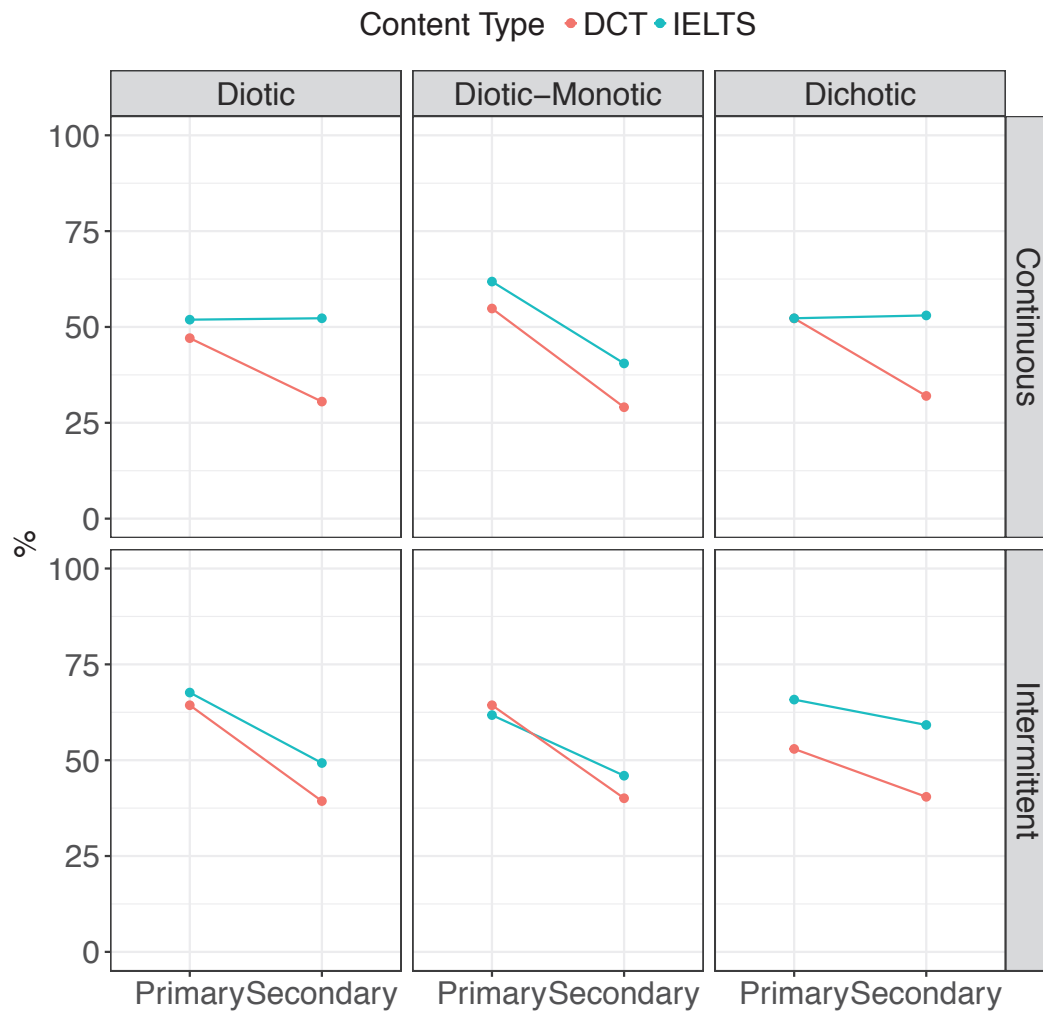


Fig. 2. Percentage of correct answers w.r.t the primary and the secondary streams in concurrent speech-based designs.

stream of IELTS. Intermittent. Dichotic design, the percentage of correct answers to the questions set from MIS remained 85% whereas in MII, DTS, DTI it remained 68%, 56%, 54% respectively. In the same IELTS. Intermittent. Dichotic design the secondary stream's percentage for MIS remained 76% whereas in MII, DTS, DTI it remained 57%, 53%, 50% respectively. This shows that the users' comprehension depth in this concurrent speech-based design remained similar to the comprehension depth calculated in the benchmark.

5.4 Users' Experience

This section briefly discusses the users' experience while interacting with the concurrent speech-based designs. Besides the questions directed on the speech content, a descriptive question, "Can you please share your experience in using this system?" was asked to understand the user's experience with the system. The users' experience, reactions, and suggestions are concisely presented to discuss the viability of concurrent speech-based communication. This provides several hints to explore the possibility of communicating speech-based information concurrently.

Table 2. Proportion comparison of correct answers between the primary and the secondary streams of designs mentioned in Table 1, with Bonferroni correction, *** <0.001

Design	p-Value
DCT	
DCT Continuous Diotic	***
DCT Continuous Diotic-Monotic	***
DCT Continuous Dichotic	***
DCT Intermittent Diotic	***
DCT Intermittent Diotic-Monotic	***
DCT Intermittent Dichotic	0.005
IELTS	
IELTS Continuous Diotic	1
IELTS Continuous Diotic-Monotic	***
IELTS Continuous Dichotic	0.932
IELTS Intermittent Diotic	***
IELTS Intermittent Diotic-Monotic	***
IELTS Intermittent Dichotic	0.132

Table 3. Results of the one-to-one proportion comparison between the correct answers in each stream of the baseline and the concurrent designs w.r.t MIS, MII, DTS, DTI, with Bonferroni correction, *** <0.001

Concurrent Design	Primary Stream				Secondary Stream			
	MIS	MII	DTS	DTI	MIS	MII	DTS	DTI
DCT								
DCT Continuous Diotic	***	0.082	0.256	0.026	***	***	0.011	0.015
DCT Continuous Diotic-Monotic	0.038	0.79	1	0.002	***	***	0.46	***
DCT Continuous Dichotic	***	0.082	0.879	0.042	***	***	0.125	0.001
DCT Intermittent Diotic	0.038	0.614	0.729	0.804	***	***	0.124	0.102
DCT Intermittent Diotic-Monotic	0.173	0.614	0.052	0.102	***	***	0.124	0.004
DCT Intermittent Dichotic	0.001	0.482	0.882	0.002	***	***	0.053	0.066
IELTS								
IELTS Continuous Diotic	0.001	0.052	0.257	0.525	0.022	0.265	0.35	0.015
IELTS Continuous Diotic-Monotic	0.001	0.634	0.579	0.292	***	0.002	0.257	0.002
IELTS Continuous Dichotic	***	0.083	0.257	0.662	0.001	0.032	0.579	0.067
IELTS Intermittent Diotic	0.173	0.785	0.346	0.065	***	0.019	0.082	0.4
IELTS Intermittent Diotic-Monotic	0.011	0.123	0.457	0.297	***	0.003	0.053	0.4
IELTS Intermittent Dichotic	1	0.625	0.586	0.804	0.173	0.051	0.882	0.961

A few users reported that it was a fairly interesting and intriguing method that carries the potential to improve the multi-tasking perspective of life, subject to better implementation and considerations of design. The concurrent approach can be useful in contexts where information isn't critical – for instance, listening about stock markets, match commentary, news reports etc. Another example could be listening to music or sounds, rather than densely layered narratives. Users also reported that they felt that use of these systems depends on an individual's mental capabilities and differing preferences. This infers that controls that allow a user to configure the precise format for how to listen to concurrent information should be given to the users so that they

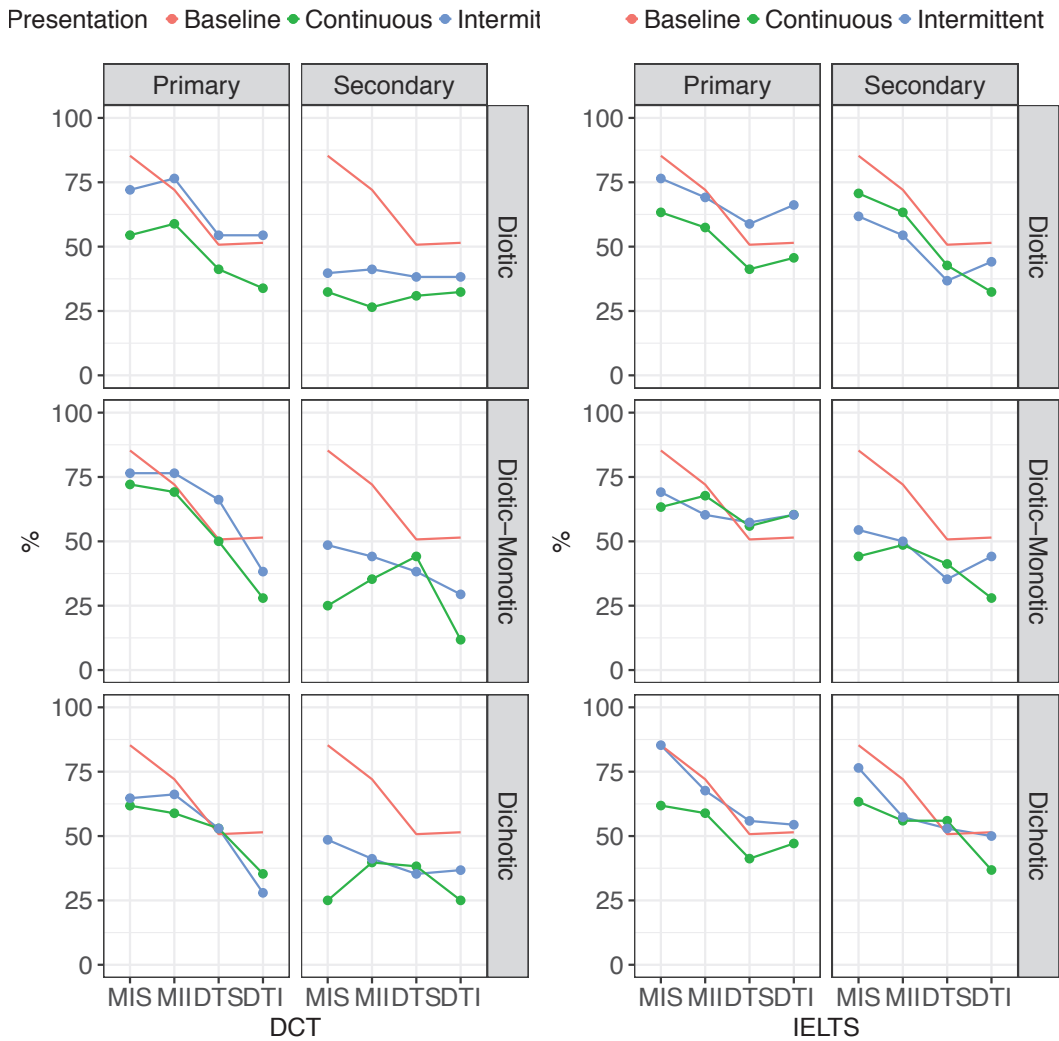


Fig. 3. The Users’ comprehension in each stream w.r.t MIS, MII, DTS, DTI for designs mentioned in Table 1. The red line is benchmark drawn from baseline condition and shown in each design for comparison.

may set up the listening session according to their preferences. Users pointed out that the female voice felt more clear and dominant as opposed to the male voice. The secondary voice presented intermittently was less challenging cognitively and was easier to comprehend, however, it created an issue of excessive attention switching. Moreover, it was reported by users that the content played with higher play-rate felt a better approach to communicating speech-based content fast as it did not affect the ability to focus on stream and remember the content. User observations regarding female voice clarity and intermittent communication proved evident in our computational analysis as the comprehension recorded was better in both the cases (section 5.2).

Some of the users completely disagreed with the idea of communicating speech-based information concurrently. They reported that the communication in parallel might mean that users miss significant information. They also reported that it was extremely difficult to focus on both the concurrent streams at the same time. The distractions made them confused and resulted in overlapping of the content from both the streams. Moreover, the identical words and phrases played together in concurrent streams made it further confusing to comprehend information. Additionally,

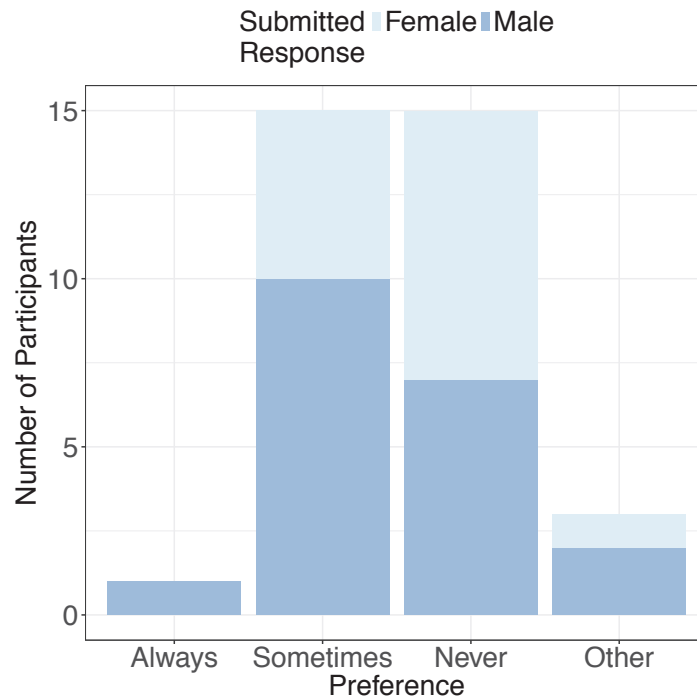


Fig. 4. Participants preference for concurrent information communication

some users reported the experiment was extremely challenging, especially when the task required both listening and answering the questions.

Finally, based on their experience in the experiment, participants were asked how frequently they would prefer concurrent communication over the baseline sequential information. Half of the users as shown in Figure 4 opted 'sometimes' whereas others said 'Never' and one said 'always' which suggest cognitive load was an issue that needs to be looked in. Regarding selecting 'sometimes' there is a need to identify the contexts where users would be looking to prefer the concurrent communication over sequential.

6 DISCUSSION

Compared to the baseline condition, user's comprehension remained significantly lower when both streams were provided continuously in speech-based concurrent design. In all the conditions where the concurrent information streams were presented continuously, the comprehension of MIS remained significantly lower than the comprehension of MIS in the baseline condition. However, the comprehension improved when additional design factors were involved in concurrent speech-based designs. For instance, the secondary information stream provided intermittently in speech-based concurrent design having stereo-channelled audio quality rendered better concurrent-speech comprehension compared to the continuous presentation. In addition to these factors, when the information was further facilitated with dichotic spatial cue, it attained similar comprehensibility that was achieved in the baseline condition. This can be seen in results (section 5.2), the comprehension remained higher in IELTS.Intermittent.Dichotic Design. In this design, the information comprehension remained similar not only in MIS but also in MII, DTS, and DTI. In conclusion, the concurrent speech-based information works better if the information is provided intermittently, dichotically and also if the audio quality of the information stream is good (stereo-channelled).

In all the intermittent and those designs where information was presented by involving Diotic-Monotic spatial difference, users comprehension remained better in the primary stream than the secondary stream. In these designs, users considered female voice (high fundamental frequency) as a primary voice to focus because of the following reasons:

Continuity The primary stream in female voice was continuous whereas the male voice was played intermittently. The continuity of stream affected the user behavior to treat the female voice as a primary voice.

Sound Pressure Level In Diotic-Monotic designs, the female stream was dominant as it was coming to both ears comparing to the male stream that was coming to right ear only. The difference in sound pressure level (SPL) contributed in treating the female voice as a primary voice to pay attention.

These findings may help to communicate two streams concurrently, one being treated as a primary voice attracting more attention of the users and the other to be treated as secondary information to provide additional information.

In this experiment, besides diotic and dichotic, another spatial cue combination, Diotic-Monotic, was introduced where the primary stream was played in both ears and the secondary stream, incited from the Right Ear Advantage (REA), was played in the right ear only. Aided by REA, it was expected that the secondary stream information would require less attention or processing for comprehension and users would be able to pay dominant attention to the primary stream played in both the ears. Consequently, this design would render better comprehensibility. The results showed that this approach didn't produce any advantage. In fact, in one of the designs, DCT.Continuous.Diotic-Monotic, the comprehension of secondary stream remained the lowest. The fundamental reason is the low intensity of the secondary stream comparing to the primary stream. Since the secondary stream in this design was coming to one ear only, therefore, the SPL of this stream was perceived lesser than the primary stream coming to both ears. However, when the SPL was same for both streams in Dichotic designs, the comprehension of secondary stream remained better than the Diotic and Diotic-Monotic designs. But in Dichotic, same SPL wasn't the only factor; the other important cue was both streams had the spatial difference of 180 degrees that also contributed in attaining better comprehension. Considering this, an interesting investigation to explore REA could be to increase the SPL of the secondary stream presented to right ear only and bring it equal to the SPL of the primary stream that is presented to both ears and then examine it. In the present experiment, the Diotic-Monotic design didn't serve any advantage in speech-based concurrent communication.

As mentioned in the method section, in DCT content the default low fundamental frequency (male) voice was increased 17% to generate an impression that the other stream is being played in a female voice. Considering the DCT.Continuous.Diotic design, where the only difference between both the streams was a difference of values in fundamental frequency, users comprehended more information from the content played in higher fundamental frequency. This result shows that the high-frequency voice attracts more attention of the listeners in case of competing voices comparing to the low-frequency voice. The application of this finding could be to use high-frequency voice in a complex sound environment to disseminate the important or critical information that requires the immediate attention of the listeners among the competing voice-based streams.

Moreover, as discussed in a method section, the questions were arranged in 4 categories MIS, MII, DTS, DTI formed on the basis of information repetition in the content to assess the comprehension depth. It was expected that the users content comprehension would remain in the same order mentioned above, as the main information was repeated multiple time in the content whereas the detailed information was played once in the content. User's comprehension remained higher in MIS

and MII and remained lower in DTS and DTI. The analysis showed that the users comprehended the main information well and were able to comprehend information to some extent i.e. about 50% in the baseline condition. In almost all the concurrent speech-based designs the percentage of correct answers remained significantly lower than the baseline condition. However, the pattern of comprehension depth remained similar in both streams played concurrently in each speech-based concurrent design as was seen in the baseline condition. User answered more questions correctly which were drawn from MIS/MI and performance remained lower in DTS and DTI. In conclusion, the pattern of comprehending information didn't change in concurrent speech-based design.

Regarding information presentation preference, 15 users said they would 'sometimes' prefer concurrent speech-based communication over sequential, and one said he would prefer it 'always'. From the users showing interest in speech-based concurrent communication, 10 were male and 5 female. This shows that male users showed a higher interest in speech-based concurrent communication than the female users.

Many users reported high cognitive load in concurrent speech-based information communication. Since the objective of this study was to assess the content comprehension by the users in concurrent speech-based communication, therefore, this experiment required users to listen content from 14 stimuli designs and answer the questions. This lengthy experiment appeared extensive and boring to the users. The experiment impacted high cognitive load and demanded extensive use of memory.

6.1 Limitations and Future Work

Though the experiment systematically assessed many factors stated in the discussion section, it does not cover how a user would practically use such systems in the real environment and what types of content in varying contexts can be played concurrently to the users. A fully functional prototype providing full control to the users to set the information flow according to their needs and context can be tested in a real environment to analyze the wild-usage. Such in-the-wild studies would help to capture usage and behaviors of the users that are not possible to capture with laboratory-based investigations [23].

7 CONCLUSION

The experiment results showed that communicating concurrent speech-based information is practical. The experimental study showed: (1) In the concurrent speech-based information communication users, besides answering the main questions, can also successfully answer some of the implied questions, as well as the questions that required detailed information. (2) The concurrent speech-based information communication works better when the information in stereo audio quality is provided intermittently and dichotically. (3) The Diotic-Monotic design involving REA doesn't serve advantage in speech-based concurrent communication. (4) The high-frequency voice attracts more attention from the listeners in competing voices. (5) Male users were more interested in speech-based concurrent communication comparing to the female users. (6) The comprehension pattern remains similar in concurrent speech-based communication as seen in sequential communication. (7) Besides encouraging results in concurrent speech-based information communication, users also reported the high cognitive load. There is a need to continue the research in the same direction, mainly addressing the high cognitive load issue by identifying the viable combinations of types of information streams and contexts where concurrent communication could be implemented.

A IELTS-BASED CONTENT

Following are a couple of Continuous and Intermittent transcripts of the IELTS-based audio streams along with the categorized questions having the answers from 'Yes | No | Don't Know' options.

A.1 Continuous Stream

A.1.1 Transcript. Welcome to Green Vale Agricultural Park. As you know, we have only been open a week so you are amongst our first visitors. We have lots of fascinating indoor and outdoor exhibits on our huge complex, spreading hundreds of hectares. Our remit is to give educational opportunities to the wider public as well as to offer research sites for a wide variety of agriculturists and other scientists. Let's start by seeing what there is to do. As you can see, here, on our giant wall plan, we are now situated in the reception block, here, as you walk out of the main door into the park, there is a path you can follow. If you give route, you will immediately come into the rare breeds section, where we keep a wide variety of animals, which I shall be telling you a little more about later. Next to this, moving east is the large grazing area, for the rare breeds.

A.1.2 Questions

. Main Information Stated (MIS)

- (1) Did the speaker describe an Agriculture Park?
- (2) Did the speaker specifically talk about how to get a bumper cotton crop?

Main Information Implied (MII)

- (3) Is the place suitable to visit by Agriculturists or Scientists?
- (4) Is the place organised into different sections?

Detailed Information Stated (DTS)

- (5) Has the place opened a month ago?
- (6) Does the place has a single variety of Animals?

Detailed Information Implied (DTI)

- (7) Is the giant wall plan situated in Reception block?
- (8) Is the rare breed section far away from the Reception block?

A.2 Intermittent Stream

In the Intermittent stream, each bullet point mentioned below was played using one chunk.

A.2.1 Transcript.

- My spoken Spanish was already pretty good in fact.
- In fact, I ended up teaching English there, although that wasn't my original choice of work.
- I found an agency that runs all kinds of voluntary projects in South America.
- Getting involved in building projects was an option. Then there was tourism - taking tourists for walks around the volcanoes.

A.2.2 Questions. Main Information Stated (MIS)

- (1) Did speaker talk about his dance skills?
 - (2) Is speaker good in the Spanish language?
- #### Main Information Implied (MII)
- (3) Is the speaker multi-lingual?
 - (4) Did user work for a role that was not of his interest?
- #### Detailed Information Stated (DTS)
- (5) Does the agency in South America runs commercial projects?
 - (6) Was it an out of options for the speaker to involve in the building projects?
- #### Detailed Information Implied (DTI)
- (7) Did speaker worked as a doctor?
 - (8) Does speaker has proficient English skills?

REFERENCES

- [1] (1997). Discourse Comprehension Test: Test KIT. http://www.picaprograms.com/discourse_comprehension_test.htm. [Online; accessed 19-October-2016].
- [2] Aydelott, J., Baer-Henney, D., Trzaskowski, M., Leech, R., and Dick, F. (2012). Sentence comprehension in competing speech: Dichotic sentence-word priming reveals hemispheric differences in auditory semantic processing. *Language and Cognitive Processes*, 27(7-8):1108–1144.
- [3] Aydelott, J., Jamaluddin, Z., and Nixon Pearce, S. (2015). Semantic processing of unattended speech in dichotic listening. *The Journal of the Acoustical Society of America*, 138(2):964–975.
- [4] Beattie, D., Baillie, L., and Halvey, M. (2015). A comparison of artificial driving sounds for automated vehicles. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 451–462. ACM.
- [5] Beattie, D., Baillie, L., and Halvey, M. (2017). Exploring how drivers perceive spatial earcons in automated vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):36.
- [6] Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening a. *The Journal of the Acoustical Society of America*, 120(3):1506–1516.
- [7] Biatov, K. and Koehler, J. (2003). An audio stream classification and optimal segmentation for multimedia applications. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 211–214. ACM.
- [8] Brayda, L., Traverso, F., Giuliani, L., Diotalevi, F., Repetto, S., Sansalone, S., Trucco, A., and Sandini, G. (2015). Spatially selective binaural hearing aids. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 957–962. ACM.
- [9] Bregman, A. (1990). Auditory scene analysis: The perceptual organization of sound. 1990.
- [10] Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.
- [11] Brungart, D. S. and Simpson, B. D. (2005). Optimizing the spatial configuration of a seven-talker speech display. *ACM Transactions on Applied Perception (TAP)*, 2(4):430–436.
- [12] Chernyshov, G., Tag, B., Chen, J., Noriyasu, V., Lukowicz, P., and Kunze, K. (2016). Wearable ambient sound display: embedding information in personal music. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, pages 58–59. ACM.
- [13] Church, K., Cherubini, M., and Oliver, N. (2014). A large-scale study of daily information needs captured in situ. *ACM Trans. Comput.-Hum. Interact.*, 21(2):10:1–10:46.
- [14] Conway, A. R., Cowan, N., and Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic bulletin & review*, 8(2):331–335.
- [15] Csapó, Á. and Wersényi, G. (2013). Overview of auditory representations in human-machine interfaces. *ACM Computing Surveys (CSUR)*, 46(2):19.
- [16] Dix, A., Finlay, J. E., Abowd, G. D., and Beale, R. (2003). Human-computer interaction.
- [17] Doherty, J., Curran, K., and McKeivitt, P. (2013). A self-similarity approach to repairing large dropouts of streamed music. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(3):20.
- [18] Elhilali, M. and Shamma, S. A. (2008). A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6):3751–3771.
- [19] Fazal, M. and Karim, M. S. (2017). Multiple information communication in voice-based interaction. In *Multimedia and Network Information Systems*, pages 101–111. Springer.
- [20] Feng, W.-c. (2012). Streaming media evolution: where to now? In *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video*, pages 57–58. ACM.
- [21] Griffiths, T. D. and Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5(11):887–892.
- [22] Guerreiro, J. (2013). Using simultaneous audio sources to speed-up blind people’s web scanning. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, page 8. ACM.
- [23] Guerreiro, J. (2016). Towards screen readers with concurrent speech: where to go next? *ACM SIGACCESS Accessibility and Computing*, (115):12–19.
- [24] Guerreiro, J. and Gonçalves, D. (2014). Text-to-speeches: evaluating the perception of concurrent speech by blind people. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 169–176. ACM.
- [25] Guerreiro, J. and Gonçalves, D. (2016). Scanning for digital content: How blind and sighted people perceive concurrent speech. *ACM Transactions on Accessible Computing (TACCESS)*, 8(1):2.
- [26] Gulliver, S. R. and Ghinea, G. (2006). Defining user perception of distributed multimedia quality. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(4):241–257.
- [27] Hinde, A. F. (2016). *Concurrency in auditory displays for connected television*. PhD thesis, University of York.
- [28] Hines, A., Gillen, E., Kelly, D., Skoglund, J., Kokaram, A., and Harte, N. (2014). Perceived audio quality for streaming stereo music. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1173–1176. ACM.

- [29] Iyer, N., Thompson, E. R., Simpson, B. D., Brungart, D., and Summers, V. (2013). Exploring auditory gist: Comprehension of two dichotic, simultaneously presented stories. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 050158. ASA.
- [30] Kortum, P. (2008). *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [31] Lawson, E. A. (1966). Decisions concerning the rejected channel. *The Quarterly journal of experimental psychology*, 18(3):260–265.
- [32] Li, G.-p. and Huang, G.-y. (2005). The "core-periphery" pattern of the globalization of electronic commerce. In *Proceedings of the 7th International Conference on Electronic Commerce, ICEC '05*, pages 66–69, New York, NY, USA. ACM.
- [33] Matassa, A., Console, L., Angelini, L., Caon, M., and Khaled, O. A. (2015). Workshop on full-body and multisensory experience in ubiquitous interaction. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, pages 923–926. ACM.
- [34] McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, 19(22):R1024–R1027.
- [35] McGookin, D. K. and Brewster, S. A. (2004). Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Transactions on Applied Perception (TAP)*, 1(2):130–155.
- [36] Moffat, D. and Reiss, J. D. (2018). Perceptual evaluation of synthesized sound effects. *ACM Transactions on Applied Perception (TAP)*, 15(2):13.
- [37] Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly journal of experimental psychology*, 11(1):56–60.
- [38] Nelson, C. (1995). Attention and memory: An integrated framework. *Oxford Psychology Series*, 26.
- [39] Obermeyer, J. A. and Edmonds, L. A. (2018). Attentive reading with constrained summarization adapted to address written discourse in people with mild aphasia. *American journal of speech-language pathology*, 27(1S):392–405.
- [40] Patel, D., Ghosh, D., and Zhao, S. (2018). Teach me fast: How to optimize online lecture video speeding for learning in less time? In *Proceedings of the Sixth International Symposium of Chinese CHI*, pages 160–163. ACM.
- [41] Qudah, B. and Sarhan, N. J. (2010). Efficient delivery of on-demand video streams to heterogeneous receivers. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 6(3):20.
- [42] Rivenez, M., Darwin, C. J., and Guillaume, A. (2006). Processing unattended speech. *The Journal of the Acoustical Society of America*, 119(6):4027–4040.
- [43] Sato, D., Zhu, S., Kobayashi, M., Takagi, H., and Asakawa, C. (2011). Sasayaki: Augmented voice web browsing experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pages 2769–2778, New York, NY, USA. ACM.
- [44] Schmandt, C. and Mullins, A. (1995). Audiostreamer: exploiting simultaneity for listening. In *Conference companion on Human factors in computing systems*, pages 218–219. ACM.
- [45] Vazquez Alvarez, Y. and Brewster, S. A. (2010). Designing spatial audio interfaces to support multiple audio streams. In *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pages 253–256. ACM.
- [46] Welland, R. J., Lubinski, R., and Higginbotham, D. J. (2002). Discourse comprehension test performance of elders with dementia of the alzheimer type. *Journal of Speech, Language, and Hearing Research*, 45(6):1175–1187.
- [47] WILLIAMS, S. M. (1994). *Perceptual principles in sound grouping*. In *Auditory Display: Sonification, Audification, and Auditory Interfaces*. Addison-Wesley.
- [48] Wu, T., Dou, W., Wu, F., Tang, S., Hu, C., and Chen, J. (2016). A deployment optimization scheme over multimedia big data for large-scale media streaming application. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(5s):73.
- [49] Xu, C., Maddage, N. C., Shao, X., and Tian, Q. (2007). Content-adaptive digital music watermarking based on music structure analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):1.
- [50] Zimmermann, R. and Liang, K. (2008). Spatialized audio streaming for networked virtual environments. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 299–308. ACM.

Received October 2018; revised October 2018; accepted October 2018

Appendix N

Publication 8 [Submitted]

——, “Investigating cognitive workload in concurrent speech-based information communication,” *The Journal of the Acoustical Society of America (JASA)*, vol. -, no. -, pp. 1–20, 2019, submitted

Investigating cognitive workload in concurrent speech-based information communication

Muhammad Abu ul Fazal,¹ Sam Ferguson,¹ and Andrew Johnston¹

School of Computer Science, Faculty of Engineering and IT,

University of Technology Sydney, NSW 2007, Australia

(Dated: 19 July 2019)

1 In conventional speech-based interaction methods, systems communicate information
2 to users sequentially, but users are actually capable of noticing, listening, and compre-
3 hending concurrent information simultaneously. This fact implies that the sequential
4 approach may be under-utilising human perception capabilities and restricting user
5 performance to sub-optimal levels. This paper reports on an experiment that inves-
6 tigates the cognitive workload experienced by the users when listening to a variety
7 of combinations of information types in concurrent formats. Fifteen different com-
8 binations of concurrent information streams were investigated, and the subjective
9 listening workload for each of the combination was measured using NASA-TLX. The
10 results showed that the perceived workload index score in concurrent and the base-
11 line condition has no significant difference. However, users response in preferring and
12 frequently using various concurrent combinations was significantly low compared to
13 the baseline condition. It is expected that the results of this experiment will help
14 digital content creators and designers to communicate information more efficiently
15 to users.

16 I. INTRODUCTION

17 In an auditory scene, users are capable of focusing their attention on speech-based infor-
18 mation streams of their choosing when they receive competing speech-based information in
19 parallel. The well-known example highlighting this phenomenon is the cocktail party prob-
20 lem ([Bee and Micheyl, 2008](#)) where a person receives multiple voice streams concurrently and
21 manages to pay attention to a particular stream using the selection and attention abilities
22 by prioritising the interest ([Cherry and Taylor, 1954](#)). In contemporary implementations
23 of voice-based interaction, the question arises: Are current information designs optimally
24 utilising human auditory capabilities for voice-based interaction?

25 Visual interfaces, the most common method of information communication, can pro-
26 vide multiple streams and sources of information in parallel to a user by using a variety of
27 methods, such as numeric displays, text, graphical representations, and even computer user
28 interface elements, one of which is the use of overlays ([Neil, 2009](#); [Ware, 2012](#)). Similar con-
29 cepts theoretically may be adopted in voice-based interaction interfaces for communicating
30 multiple information concurrently, because the human auditory system is capable of per-
31 forming filtering of the sounds received and can allow users to ignore extraneous noise and
32 concentrate on relevant information ([Bregman, 1994](#); [Dix, 2003](#)). For example, a possible
33 increase in the information communicated could theoretically be provided by broadcasting
34 two voice streams concurrently, one as a primary stream representing the main information,
35 and the other stream, as an assistant that provides additional information based on the
36 context and behaviour ([Sato et al., 2011](#)). However, designing such concurrent information

37 streams can be a real challenge that would decide whether such communication method is
38 helpful to the users, or distracts users in interacting with the system.

39 There are many other applications of concurrent information communication ([Guerreiro,](#)
40 [2016](#)). For many applications, such as concurrent speech synthesisers, Interactive Voice
41 Response systems (IVRs), and obtaining audio or video information within large corpi,
42 listening to two concurrent streams and gaining a gist from multiple information streams
43 concurrently could be of great utility. For example, a user might listen to various live talk
44 shows that focus on different topics. That user might be interested in listening to more than
45 one live program at the same time, such as a talk show discussing politics while listening
46 to a program that discusses music. A few other activities among the wider population that
47 motivate to explore concurrent communication are:

- 48 1. Students engaged in study may have multiple screens at hand. While studying, the
49 student might have their laptop that they are working on, their phone in arms reach
50 and a television/radio/stream playing in the background.
- 51 2. Parents who are obliged to have a child's program playing on a large television screen,
52 while they have their programming on a secondary (possibly less audible) screen.
53 The parent is likely to be attempting to pay attention to both streams to ensure
54 that appropriate content is playing on the television screen for the child while being
55 entertained by their programming choice.
- 56 3. Video game players may have an instructional video streaming on one screen while
57 they are gaming on a second screen.

58 Besides the less critical applications of concurrent speech-based information communi-
59 cation, many other critical real-life domains may benefit from a systematic understanding
60 of concurrent information communication designs. Professionals who engage in listening
61 to multiple talkers simultaneously, such as air traffic controllers, watchstanding sailors, or
62 physicians working in an emergency ward, who generally balance competing priorities, du-
63 ties, and tasks by listening and interacting with multiple sources simultaneously (Walter
64 *et al.*, 2017) may benefit from concurrent designs. In the medical industry, research is
65 already heading where auditory displays enable the head-up monitoring of the patient dur-
66 ing theatre operations (Sanderson, 2006). Similarly, possibilities of non-speech concurrent
67 communication have also been explored in the context of flight-decks (Towers, 2016). Con-
68 current speech-based information communication in such critical fields would require careful
69 considerations and research.

70 Exploring concurrent information communication is generally important as it may enable
71 users to perform roles and tasks efficiently, and therefore research interest in this topic
72 has recently been increasing. As discussed in the section-II, researchers have delineated
73 many parameters that could improve concurrent information communication to users. The
74 research also has shown that concurrent information communication creates a high amount
75 of cognitive challenge for users when listening to multiple information streams (Fazal *et al.*,
76 2018a; 2019; Xia *et al.*, 2015). In this regard, to the best of authors' knowledge, what has
77 not been explored is which types of information streams can be concurrently listened to
78 without creating excessive cognitive load. This research explores concurrent speech-based
79 information communication where various types of information streams (including musical

80 streams) are combined to comprehensively investigate the cognitive workload encountered
81 when listening to concurrent information streams.

82 The paper is organised as follows: Section II presents Background, Section III presents
83 Aims & Motivation, Section IV presents Methodology, Section V presents Results, Section VI
84 presents Discussion, and Section VII presents Limitations and Future Work.

85 II. BACKGROUND

86 Many researchers (Fazal *et al.*, 2018b; Fazal and Shuaib Karim, 2017; Feltham and Loke,
87 2017; Guerreiro and Goncalves, 2016; Hinde, 2016; Ikei *et al.*, 2006; Mullins, 1996; Parente,
88 2008; Schmandt and Mullins, 1995; Towers, 2016; Werner *et al.*, 2015) have worked on intro-
89 ducing concurrent communication through auditory display, and have reported remarkable
90 performance by participants listening to two simultaneous voice streams, showing that a
91 listener can process secondary information present in the voice stream that is not the imme-
92 diate focus. For instance, *AudioStreamer* by Schmandt and Mullins (1995) is one of the first
93 speech interfaces that endeavoured to use people’s ability to attend the desired stream from
94 the competing streams selectively. In this system, three concurrent speech-based streams
95 were presented, applying spatial manipulation to each. The system was designed to track
96 head movement to identify the user’s interest in a stream of the competing streams. Mullins
97 (1996) stated that *AudioStreamer* users were cognitively overwhelmed by three channels
98 of concurrent speech. To overcome this, Mullins suggested introducing five-second onset
99 asynchronies between the streams. Schmandt (1998) introduced *Audio Hallway* as his sec-
100 ond speech interface exploiting the concurrent speech-based presentation that allowed the

101 browsing of vast compilations of audio files. [Parente \(2008\)](#) developed a speech interface
102 prototype, called Clique, where users instead of interacting with the underlying graphical
103 interfaces, listened to and interacted solely with the display. For improved television ex-
104 periences, [Hinde \(2016\)](#) explored how auditory displays can offer an alternative method
105 that depends on users' desire to being able to attend screen-based information. The results
106 showed that offering sound-based secondary content from a smartphone after removing the
107 speech from the television program was the best auditory approach. There are many pro-
108 totypes addressing user's interaction with the system are introduced by the researchers to
109 communicate speech-based information concurrently.

110 As the concurrent information communication may also aid in critical domains, the re-
111 searchers in the U.S. Naval Research Laboratory (NRL) for improving the Navy watch
112 standing operations conducted a studies ([Brock *et al.*, 2008; 2011](#)) aimed at developing a
113 set of comparative measures of attention and comprehension in a variety of multi-talkers
114 information contexts involving concurrent and serial speech communications. Similarly, for
115 improving a pilots situational awareness for the changing state of systems information, [Tow-](#)
116 [ers \(2016\)](#) supported the use of spatial auditory displays within flight decks. The results
117 of the studies supported the use of concurrent spatial sonifications as it helped users to
118 spend more head-up time to an out of flight deck visual search task and fly the aircraft more
119 precisely.

120 The speech-based interaction also benefits visually impaired persons for interacting with
121 the system as visually impaired users mostly rely on their auditory system to receive infor-
122 mation. [Guerreiro and Goncalves \(2016\)](#) carried out research on blind and sighted users, and

123 conducted experiments to determine the information scanning abilities of the sighted and
124 the visually impaired person from the concurrent speech. [Guerreiro and Goncalves](#) lever-
125 aged the concept of cocktail party problem. [Guerreiro and Goncalves \(2016\)](#) found that
126 the spatial difference in sources is the best cue in concurrent speech. The study established
127 that sighted and the visually impaired users have similar abilities to scan the information
128 from the concurrent speech ([Guerreiro and Goncalves, 2016](#)). Two concurrent information
129 streams were more useful in understanding and identifying the content. The study showed
130 that the use of three speech sources depends on the task intelligibility demands and lis-
131 tener capabilities. In another study by [Guerreiro \(2013\)](#), it was found that the concurrent
132 speech with slightly higher playback-rate enables a significantly quicker scanning for relevant
133 content. [ul Fazal et al. \(2019\)](#) based on their study involving blind and sighted users pro-
134 posed *Vinfomize* framework that may help in developing systems to communicate multiple
135 voice-based information to the users subject to users' contextual and perceptual needs and
136 limitations.

137 For providing guidelines to designers to build concurrent speech interfaces, [Fazal et al.](#)
138 ([2019](#), [2018b](#)) reported on an experiment that aimed to test different speech-based designs
139 for concurrent information communication. Two audio streams from two types of content
140 were played concurrently to 34 users, in both continuous or intermittent form, with the ma-
141 nipulation of a variety of spatial configurations (i.e. Diotic, Diotic-Monotic, and Dichotic).
142 In total, 12 concurrent speech-based design configurations were tested with each user. The
143 results showed that the concurrent speech-based information designs involving intermittent
144 form and the spatial difference in information streams produce comprehensibility equal to

145 the level achieved in sequential information communication. Many users reported high cog-
146 nitive load in concurrent speech-based information communication. [Vazquez Alvarez and](#)
147 [Brewster \(2010\)](#) used a divided-attention task and conducted an experiment where an audio
148 menu and continuous podcast competed for attention. In the experiment, the impact of
149 the cognitive load was assessed using the NASA-TLX subjective assessment tool. The re-
150 sults showed that users' ability to attend two concurrent streams enhances by spatial audio,
151 and also the divided attention creates cognitive load and impacts the overall performance
152 significantly.

153 This paper extends exploring concurrent speech-based information communication and
154 comprehensively investigates the cognitive workload experienced when listening to a variety
155 of combinations of information types in concurrent formats.

156 III. AIMS & MOTIVATION

157 A. Aims

158 This experiment aims to comprehensively analyse the cognitive load by subjectively mea-
159 suring workload that users endure while listening to two different audio streams concurrently.
160 We sought to obtain data to satisfy the following questions: a) Does the cognitive workload
161 remain similar in each concurrent combinations? b) Irrespective of combinations, which in-
162 formation type(s) are preferred most by users when presented in concurrent combinations?
163 c) Do users presented with different combinations show differences in preference and likely
164 frequency of use? d) Does an intermittent form of communication in one of two streams cre-

165 ate lower cognitive workload in speech-based information communication, when compared
166 to the two continuous concurrent streams?

167 **B. Motivation**

168 The motivation behind conducting this experiment is to determine the viability of com-
169 municating concurrent information within a scenario likely to be encountered by prospective
170 users. Since the study involves dichotic presentation of songs and non-vocal music with other
171 information types not having background music, it is hoped that the study will smooth a
172 path to deliver information more quickly. Additionally, some forms of information (for in-
173 stance, headlines, tweets, or RSS feeds) an intermittent design may be a natural choice, and
174 may also help with lowering cognitive load, and this study will attempt to ascertain whether
175 the intermittent form decreases cognitive load. Knowledge from these studies may help to
176 inform the design auditory overlays that are similar to visual overlays used in many visual
177 user interface approaches as mentioned in the introductory section I.

178 **IV. METHOD**

179 The method adopted for this experiment is outlined below.

180 **A. Participants**

181 After receiving institutional Human Research Ethics Committee approval for the research
182 protocol, user participation campaigns were launched. The participants were selected based

183 on two criteria: 1) not having a significant hearing impairment, and 2) having competent
 184 English language skills, as the listening experiment’s content was in the English language.
 185 The users, selected for participation, were offered gift cards worth 30 AU\$ each. In total, 40
 186 participants, 20 female, and 20 male took part in the experiment after providing consent.
 187 The mean age of the participants was 23 with the standard deviation of 6.

TABLE I. **Combinations of Different Types:** of Information Streams in the Concurrent Stimuli

	Monolog	Inter.	Comm.	News	Songs	Music
Right Ear	Left Ear					
Monolog	-	-	-	-	-	-
Interview	✓	-	-	-	-	-
Commentary	✓	✓	-	-	-	-
News	✓	✓	✓	-	-	-
Songs	✓	✓	✓	✓	-	-
Music	✓	✓	✓	✓	✓	-

188 B. Design

189 1. *Concurrent Condition*

190 In this condition, two concurrent information streams were communicated in a dichotic
 191 form to users. For concurrent communication, we created a series of stimuli where each

192 stimulus was created by combining the two different types of information streams. One
 193 stream was presented in the right ear, and the other stream was presented in the left ear,
 194 using the panning feature in an open source software Audacity ([Audacity](#)).

195 The concurrently presented combinations included two of the following sources: Monolog
 196 (Documentary), Interview (Dialog), Commentary (Football), News Headlines, Song (Vocal),
 197 Music (Non-Vocal), representing a range of expected types of auditory content a user of a
 198 auditory presentation system might encounter (eg. a mobile phone user or computer user).
 199 Each type of information stream was combined with the other types of information streams
 200 once, as described in Table I and illustrated in Figure 1.

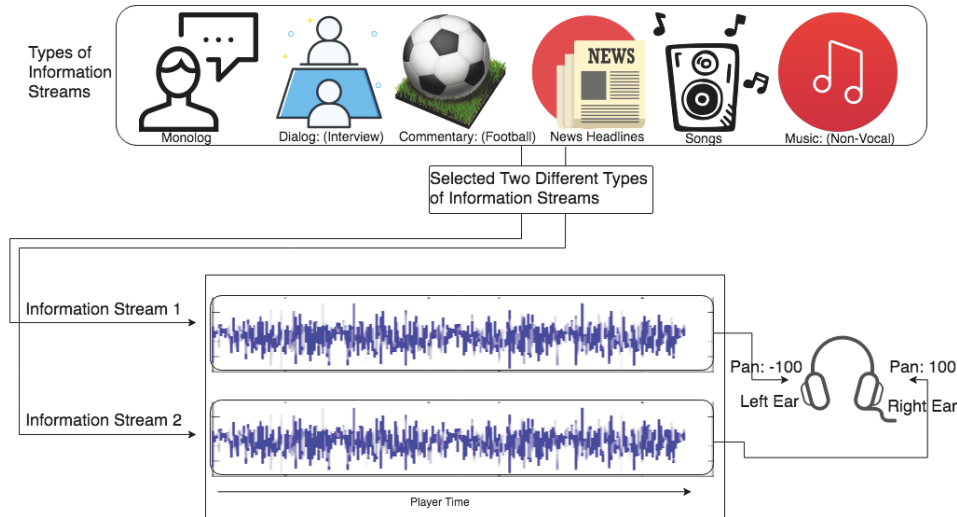


FIG. 1. Concurrent Stimulus Design

201 *a. Intermittent.* In all the stimuli, the information streams were presented continu-
 202 ously, except for the stimuli where an information stream was combined with the news
 203 headlines information stream. Following the intermittent design introduced in our previous
 204 study ([Fazal et al., 2018b](#)), we manipulated the news headlines stream and transformed it

205 to intermittent form from the continuous information presentation. For this, the news head-
206 lines bulletin was broken into temporal segments, and after each news headline, a gap of 15
207 seconds of silence was added. We involved this intermittent form as in our previous studies
208 it was found that the users' comprehension was the best in concurrent information com-
209 munication formats. Therefore, the information streams combined with the news headlines
210 information stream was of an intermittent concurrent design type. The rest of the stimuli
211 were based on continuous information design.

212 *2. Baseline Condition*

213 In this condition, no concurrency was involved. A type of information stream was ran-
214 domly selected and sequentially communicated to the users in a conventional form. The
215 purpose of this condition was to set a users' experience benchmark, and later use it to
216 compare the experience with concurrent conditions, where two information streams were
217 presented concurrently.

218 **C. Material**

219 First, the streams of six information types were selected that included: Monolog (doc-
220 umentary), Dialog (interview), Commentary (football), News Headlines, Songs (vocal) and
221 Music (instrumental / non-vocal). For each information type, BBC online channels were
222 searched to find high quality samples of information presentations. For each information
223 type, six samples of a maximum of 2 minutes duration were selected, except for the com-
224 mentary information type. For commentary, the first twelve minutes of the sports match

225 was broken in 6 equal (in duration) files. Following the selection method, there were 36 files
226 in total, each with a duration of 2 minutes.

227 Based on the ecological choices, the *monolog* streams were selected from the BBC pro-
228 gram *Lip Service* wherein each selected documentary a woman discussed a trait of her life.
229 Interviews (*dialog*) were selected from the BBC's program *BBC Celebrity Interview* where
230 a male host interviewed a male celebrity. The sports *commentary* was from the BBC 5 Live
231 Radio and was recorded in the male voice covering a football match between Napoli and
232 Manchester City. The *news headlines* spanned six different dates and were selected from
233 the BBC World News. Three news headlines were recorded in the female voice and three in
234 the male voice. The *songs* were selected from the BBC Radio-1 Channel where three of the
235 singers were female and three male. For the *music*, the background music of the Hollywood
236 movie *Viceroy* composed by the Academy Award winner AR Rahman was selected.

237 **D. Stimuli Presentation**

238 Since each type of information stream was combined with the rest of the types of infor-
239 mation streams, users were presented with all 15 different concurrent combinations. Besides
240 the concurrent combinations, a baseline stimulus was also presented. Hence, a user was
241 presented with the 16 different stimuli (15 Concurrent, 1 Sequential).

242 In total, 576 stimuli combinations were created, including the stimuli representing the
243 baseline condition in order to remove the combinational effect. A user was presented with 15
244 stimuli, each representing one combination, as well as 1 that was the baseline stimulus. In this
245 randomisation, the combinational effect was removed to make sure users were not provided

246 information streams that repeated information types. The order of presenting combinations
247 was random to remove the ordering effect.

248 E. Measures

249 The duration of each stimulus was 2 minutes. After listening to each stimulus, users
250 were presented with a questionnaire, attached as Appendix VII to share their experience.
251 The experience was obtained using the NASA-TLX subjective, multidimensional assessment
252 tool (Hart and Stavenland, 1988; NASA, 2018b). NASA-TLX is a standardised instrument
253 that rates perceived workload in order to assess a task system, or a team's effectiveness or
254 other aspects of performance. Besides being cited in over 4400 research studies, this tool
255 has been used in many research studies investigating concurrent communication (Hinde,
256 2016; Parente, 2008; Towers, 2016; Truschin *et al.*, 2014; Vazquez-Alvarez *et al.*, 2015, 2014;
257 Vazquez Alvarez and Brewster, 2010). The test has two parts. In the first part, the total
258 workload is measured using the following NASA (2018a) subjective subscales:

- 259 1. "Mental Demand - How mentally demanding was the task?
- 260 2. Physical Demand - How physically demanding was the task?
- 261 3. Temporal Demand - How hurried or rushed was the pace of the task?
- 262 4. Overall Performance - How successful were you in accomplishing what you were asked
263 to do?
- 264 5. Effort - How hard did you have to work to accomplish your level of performance?

265 6. Frustration Level - How insecure, discouraged, irritated, stressed, and annoyed were
266 you?”

267 The second part of the NASA-TLX procedure intends to create an individual weighting
268 of the above mentioned six subscales by asking the subjects to compare the dimensions
269 in a pairwise manner, based on their perceived importance. After this, some arithmetic
270 operations are used to compute the perceived workload index, which is a value from 0 to
271 100.

272 In order to gain more information about the user experience, we added an additional two
273 questions:

274 7. Like (Preference) - How much did you like this combination ?

275 8. Frequent - How frequently will you be using this combination of streams?

276 **F. Apparatus**

277 In order to minimise participation time, a web-based system using PHP, MySQL, JQuery,
278 HTML5, CSS, and Bootstrap was developed to play the stimuli. Sixteen HTML audio
279 players were designed to play each stimulus design that was presented in sequential web
280 pages. Users were only able to move to the next stimulus, when they submitted their
281 NASA-TLX form response for the current stimulus. Users responses were directly recorded
282 into a database.

283 The tests were conducted in a quiet purpose-built room in the Creativity and Cognition
284 Studios (CCS) of the University of Technology, Sydney. Three identical Apple iMac comput-

285 ers, having 2.7GHz quad-core Intel Core i5 processor, 8GB RAM, installed with Yosemite
286 10.10.5 OS were arranged in the studio. To listen to the audio stimuli Beyerdynamic's
287 DT770 250 OHM headphones were used that were connected to the headphone jack of the
288 computer. Users were provided with control of the gain of the system, in order to find a
289 comfortable listening level. Since the three computers were used in the studio, up to three
290 participants were engaged in the study simultaneously.

291 **G. General Procedure**

292 The selected users were verbally briefed on the study protocol before the start of the
293 study. Instructions were presented on a screen after registration. Before starting the study,
294 users entered their demographic profile information that included, name, email, age, gender,
295 primary language, qualification, profession, country, mood and hearing/visual impairment
296 status (although all responses were optional). At the end of the study, users' detailed
297 responses relating to their experience of information communication. User responses were
298 recorded in a database.

299 The entire experiment interface including form obtaining user's demographic profile in-
300 formation, playable URLs of stimuli representing each combination, and NASA-TLX based
301 questionnaire is produced in Appendix VII.

302 **V. RESULTS**

303 There were four approaches we undertook to analyse the experimental results: a) the
304 baseline condition was compared to the overall concurrent communication condition. b) the

305 analysis was extended by determining the subjective workload index for each combination,
306 as compared to the baseline condition. c) the impact of each information stream type on
307 the user's experience when combined with the rest of the information stream types was
308 explored. d) we determined the workload index and other experiential observations of the
309 information stream types with respect to their presentation in the left ear and the right ear.
310 These analysis steps are discussed in the subsequent subsections.

311 **A. Baseline vs. Concurrent**

312 We commenced with an analysis on the baseline condition and calculated the mean ratings
313 for each subscale, along with determining the NASA-TLX workload index. After this, the
314 same procedure was performed with the concurrent communication as a whole, without
315 taking the combination types into account and, finally, a comparison was completed between
316 the concurrent and the baseline condition.

317 **1. Baseline Condition**

318 The baseline mean scores for each rating scale is shown in Figure 2. The mean rating for
319 the mental demand in baseline condition was 36.75, whereas, for physical demand, temporal
320 demand, effort, frustration and performance was 28.25, 30.00, 38.00, 24.50, 84.12 respec-
321 tively. Using these subjective subscale ratings, combined with the weighting measure of
322 the NASA-TLX, the calculated mean index score for baseline listening task appeared 50.81.
323 Similarly, regarding the frequent scale, that is asking users how frequently they would lis-
324 ten to the baseline condition, the mean rating was 61.37. For the baseline condition, the

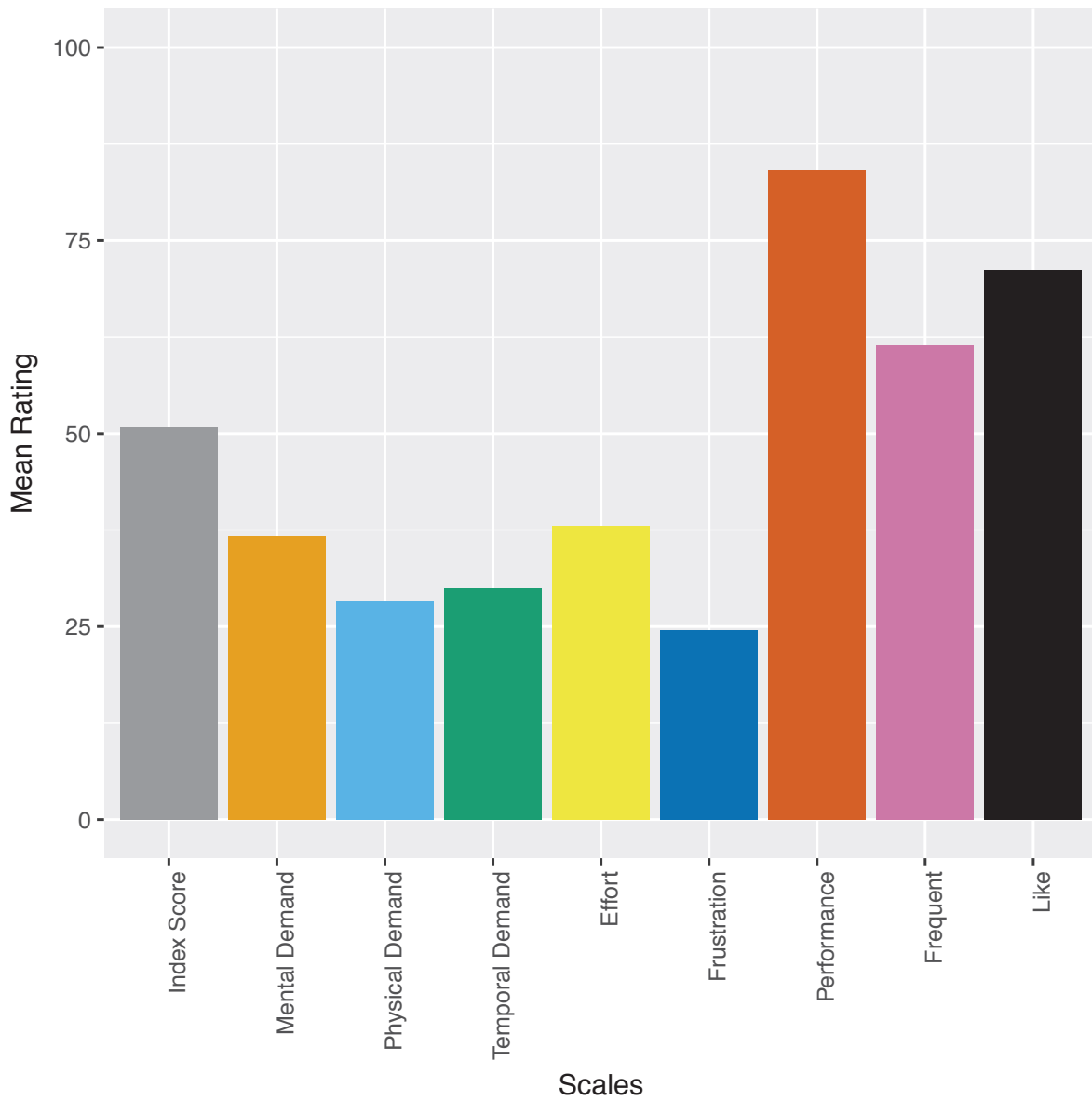


FIG. 2. Experience in Baseline Condition: shows perceived workload index score for listening task and ratings for subscales, including the frequent and preference (like) scales.

325 mean preference rating was 71.12%. These ratings set a benchmark that was used to draw
 326 comparisons with the concurrent combinations.

327 **2. Concurrent Communication**

328 Following the pattern adopted in the baseline condition, the mean score was calculated for
329 each scale, as illustrated in Figure 3. As shown in the Figure 3, the mean rating for the mental
330 demand in baseline condition was 55.24, whereas for physical demand, temporal demand,
331 effort, frustration, and performance was 40.87, 46.27, 56.72, 42.92, 62.82 respectively. Using
332 these subjective subscale ratings combined with the weighting measure of the NASA-TLX
333 the calculated mean index score for concurrent listening task appeared 59.12. Similarly,
334 regarding how frequently users would listen to the concurrent condition, the mean score was
335 37.06, and the mean score for their preference of the concurrent condition was 42.07.

336 To statistically compare the mean concurrent ratings with the baseline condition, we used
337 two-way *analysis of variance* (ANOVA) test (Copenhaver and Holland, 1988). The ANOVA
338 results, mentioned in table II, showed that the presentation type (baseline — concurrent
339 condition) does not have a significant impact on user response, $F(1, 5742) = 2.359, p <$
340 0.125 . However, the interaction between the presentation and rating scales, $F(8, 5742) =$
341 $27.098, p < 0.01$, had a significant impact on user response.

342 Since the interaction between the presentation and rating scales was significant, we per-
343 formed the *Post hoc* Tukey HSD analysis (Miller, 198; Yandell, 1997) to compare each
344 concurrent mean rating with the baseline mean rating. The results showed that the index
345 score regarding the baseline condition and the concurrent condition does not have a signifi-
346 cant difference. However, all rating scales, except physical, appeared significantly different
347 in the baseline condition and the concurrent condition ($p < .05$). Users' responses showed

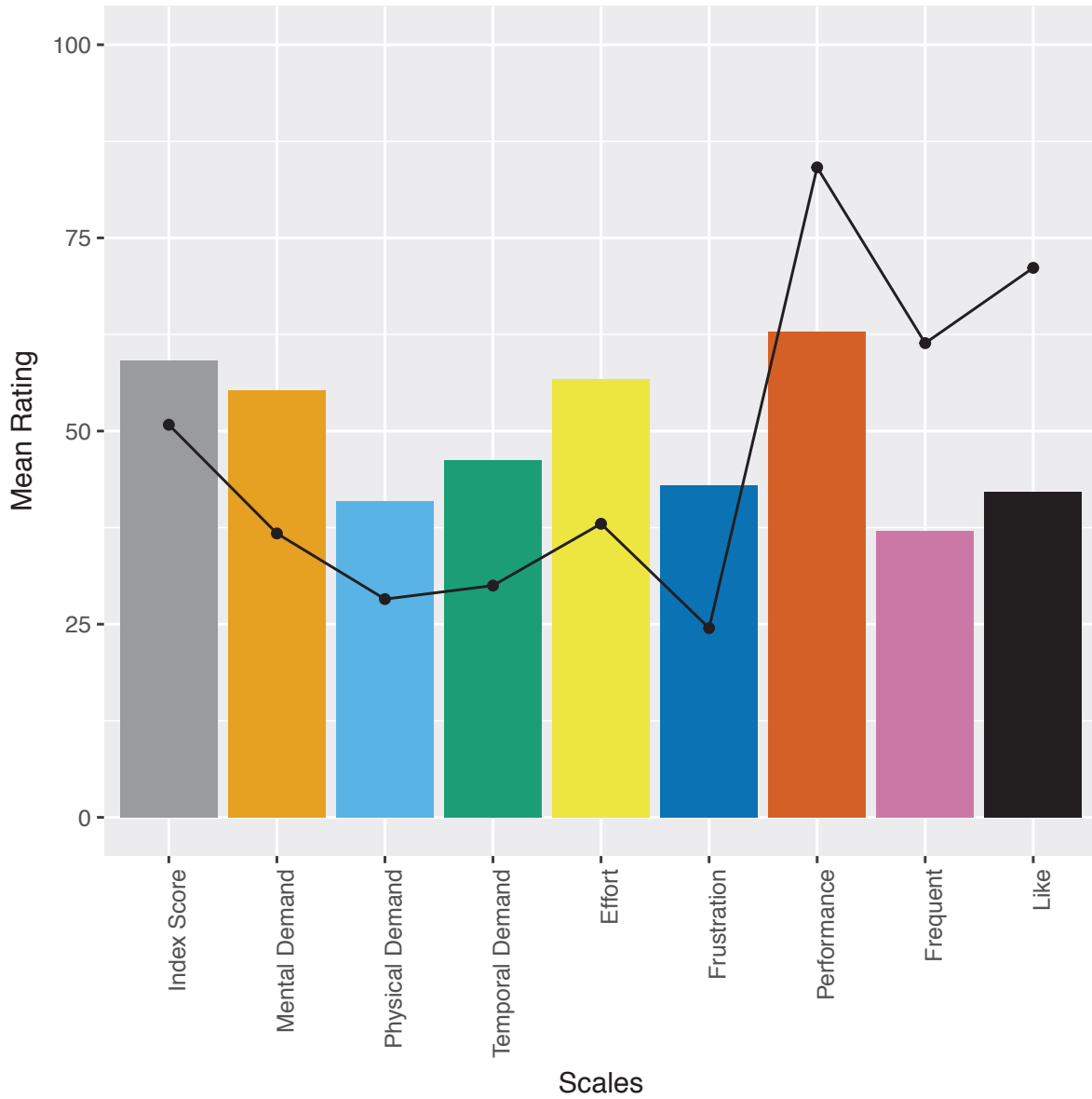


FIG. 3. Users' Experience in Concurrent Communication: compared with the baseline condition that is shown with the black continuous line.

348 that they preferred the baseline condition, and, therefore, would more frequently use it when
 349 compared to the concurrent condition. Table II shows the statistical difference between the
 350 concurrent condition and the baseline condition for each rating scale.

TABLE II. Post hoc Tukey HSD Analysis: (p -values) comparing mean ratings of concurrent scales with the relevant baseline scales: (*Signif.codes* : $\leq 0.001 = ***, \leq 0.01 = **, \leq 0.05 = *$)

Index	Ment.	Phys.	Temp.	Effo.	Frust.	Perf.	Freq.	Like
0.81	***	0.12	**	***	***	***	***	***

352 B. Concurrent Combinations

353 The results of each combination type have been compared with the baseline condition.
 354 The mean scores of the scales for all the combination designs are individually illustrated in
 355 Figure 4. The comparison of each of the mean scores with the baseline condition is also
 356 depicted using a continuous black line indicating the mean values in the baseline condition.

357 Besides illustrating the results in Figure 4, we discuss the concurrent combinations with
 358 the ANOVA results descriptively in following subsections.

359 To discuss the combination results, we categorised combinations into two types: 1)
 360 Speech-based information combinations, 2) Music-based (vocal and instrumental) experi-
 361 ence combinations. In the speech-based information combinations type, the combinations
 362 having both the streams from speech-based information types (that is, monolog, interview,
 363 commentary, news headlines) were categorised, whereas, in music-based experience combina-
 364 tions, the combinations having a stream either from the song or music types were categorised.

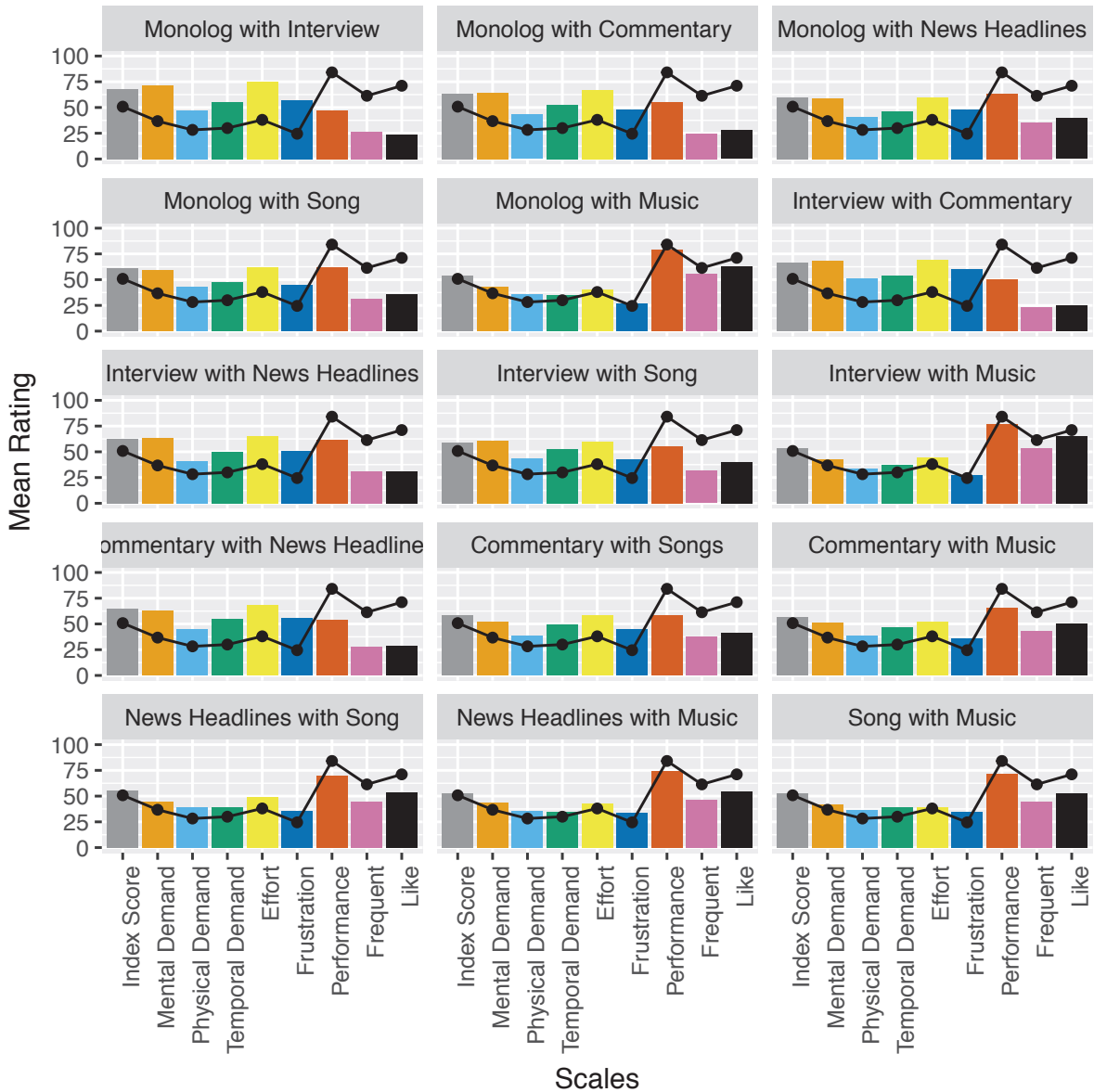


FIG. 4. Users' Experience in each Combination: of concurrent communication, also compared with the baseline condition shown with a continuous line.

365 **1. Speech-based information combinations**

366 In this category, we discuss monolog with interview, monolog with commentary, monolog
 367 with news headlines, interview with commentary, interview with new headlines, and com-

mentary with news headlines combinations. For all of these combinations, following the analysis approach we adopted in the baseline and the concurrent condition, we first calculated the mean of the responses submitted by the users against rating subscales in each combination, and then based on the means of rating subscales and the weights, we calculated the perceived workload index score for listening task in each combination.

In monolog with interview combination, the mean values of the mental demand, physical demand, temporal demand, effort, frustration, and performance were 71.75, 47.25, 55.50, 75.00, 57.00, 46.87 respectively, whereas, the index score for this combination was 67.74. Similarly, regarding the frequent and preference scales of this combination, the mean ratings was 26.62, and 23.75 respectively. Two-way ANOVA comparing the mean scores of the monolog with interview combination with the baseline condition showed that the presentation type (monolog with interview — baseline) has a significant impact on user response, $F(1, 702) = 9.71, p < 0.002$. Also, the interaction between the presentation and rating scales, $F(8, 702) = 48.125, p < 0.01$, had a significant impact on user's response. The *Post hoc* Tukey HSD test on the ANOVA results of the interaction between the presentation types and rating scales showed significant differences ($p < 0.05$) in all the scales, except, the mean index score ($p = 0.07$). Table III shows the statistical difference ($p - values$) between this combination and the baseline condition concerning each rating scale. The results showed that users' experience in this combination was not as 'good' as in the baseline condition.

In the monolog with commentary combination, the mean index score for the listening task was 63.19. The comparison of this combination with the baseline condition conducted with ANOVA showed that the presentation type (monolog with commentary — baseline) does

TABLE III. *Post hoc* Tukey HSD Analysis: (p - values) comparing mean ratings of speech-based concurrent combinations scales with the relevant baseline scales: (*Signif.codes* : $\leq 0.001 = **$ *, $\leq 0.01 = **$, $\leq 0.05 = *$)

Combination	Ind.	Men.	Phy.	Tem.	Eff	Fru.	Per.	Fre.	Lik.
Mon. w. Int.	0.07	***	*	***	***	***	***	***	***
Mon. w. Com.	0.63	***	0.29	**	***	***	***	***	***
Mon. w. New.	0.97	**	0.58	0.17	**	***	**	***	***
Int. w Com.	0.18	***	**	**	***	***	***	***	***
Int. w New.	0.70	***	0.64	*	***	***	**	***	***
Com. w New.	0.45	***	0.12	***	***	***	***	***	***

390 not have a significant impact on user response, $F(1, 702) = 1.815, p < 0.178$. However, the
391 interaction between the presentation and rating scales, $F(8, 702) = 32.44, p < 0.01$, had a
392 significant impact on the user's response. As seen in the monolog with interview combination
393 comparison, the *Post hoc* Tukey HSD test showed significant differences ($p < 0.05$) in almost
394 all scales, except, the mean index score and the physical rating scale ($p > 0.05$). Similar to
395 the monolog with interview combination, the results in this combination showed that users'
396 experience was not as 'good' as noted in the baseline condition.

397 As mentioned in the methodology section, the new headlines in all combinations were pre-
398 sented intermittently with other continuous streams concurrently to the users. In monolog

399 with news headlines, the mean index score for this listening task was 59.62. The ANOVA
 400 test on this combination showed that the presentation type (monolog with news headlines —
 401 baseline) does not have a significant impact on user response, $F(1, 702) = 3.036, p < 0.082$.
 402 However, the interaction between the presentation and rating scales, $F(8, 702) = 18.567, p <$
 403 0.01 , had a significant impact on the user's response. In the extended analysis using *Post*
 404 *hoc* Tukey HSD test performing the comparison between this combination and the baseline
 405 condition, no significant difference appeared in index score ($p = 0.97$). However, probably
 406 because of the significant difference in other rating subscales ($p < 0.05$), users significantly
 407 preferred ($p < 0.05$) the baseline condition for frequent use and preference over monolog
 408 with news headlines combination.

409 In the interview with commentary combination, the mean index score for the listening
 410 task was 66.65. The statistical analysis ANOVA showed that the presentation type (inter-
 411 view with commentary — baseline) has a significant impact on user response, $F(1, 702) =$
 412 $6.749, p < 0.01$. Also, the interaction between the presentation and rating scales had a sig-
 413 nificant impact on the user's response, $F(8, 702) = 42.724, p < 0.01$. As seen in the monolog
 414 with commentary combination comparison, the *Post hoc* Tukey HSD test showed significant
 415 differences ($p < 0.05$) in all the scales, except, the mean index score. Similar to monolog
 416 with commentary combination, the results in this combination showed that users' experience
 417 was not as 'good' as noted in the baseline condition.

418 In the interview with news headlines combination, where news headlines were presented
 419 intermittently with a continuous interview stream, the mean index score for listening this
 420 combination was 62.89. The ANOVA test on this combination showed that the presentation

421 type (interview with news headlines — baseline) has a significant impact on user response,
422 $F(1, 702) = 3.944, p < 0.047$. Also, the interaction between the presentation and rating
423 scales had a significant impact on the user's response, $F(8, 702) = 25.729, p < 0.01$. In the
424 extended analysis using *Post hoc* Tukey HSD test performing the comparison between this
425 combination and the baseline condition, significant difference ($p < 0.05$) appeared in all the
426 rating scales, except, the index score, and physical demand. The analysis shows that the
427 users found this combination a challenging experience, and therefore, significantly preferred
428 ($p < 0.05$) the baseline condition for frequent use and preference despite being provided with
429 intermittent news headlines.

430 In the commentary with news headline combination, the commentary was presented with
431 intermittent news headlines. We used commentary in combination designs on the assumption
432 that users usually do not pay in-depth attention to commentary types of information streams.
433 Users mostly remain interested in gaining the gist from the match that they can get on
434 different points, for example, the commentator becomes louder and passionate indicating
435 that something interesting is happening on the field. Such cues may help the users to divert
436 their attention immediately towards commentary with increased focus, else, pay attention
437 towards the concurrent streams. In the commentary with news headline combination design,
438 the mean index score for listening task was 64.21. The ANOVA test on this combination
439 showed that the presentation type (commentary with news headlines — baseline) has a
440 significant impact on the user response, $F(1, 702) = 4.711, p < 0.03$. Also, the interaction
441 between the presentation and rating scales had a significant impact on the user's response,
442 $F(8, 702) = 34.647, p < 0.01$. In *Post hoc* Tukey HSD test performing the comparison

443 between this combination and the baseline condition, a significant difference ($p < 0.05$)
 444 appeared in almost all the rating scales, except, the index score, and physical demand scales
 445 ($p > 0.05$).

446 2. *Music-based (Vocal and Instrumental) experience combinations*

447 In addition to two concurrent speech-based information stream combinations, song (vocal)
 448 and instrumental (non-vocal) music streams in different combinations were included on the
 449 assumption that they would enhance user experience.

450 In the monolog with song combination, a song was presented with a monolog stream
 451 and the user experience was evaluated. In this combination, the mean index score for the
 452 listening task was 60.88. The two-way ANOVA test comparing this combination with the
 453 baseline condition showed no significant difference in users response concerning presentation
 454 type (monolog with song — baseline), $F(1, 702) = 1.571, p < 0.21$. However, the interaction
 455 between the presentation and rating scales had a significant impact on the user’s response,
 456 $F(8, 702) = 21.979, p < 0.01$. In the extended analysis using *Post hoc* Tukey HSD test per-
 457 forming the comparison between this combination and the baseline condition, no significant
 458 difference ($p > 0.05$) appeared in index score. However, frequent, and like scales were sig-
 459 nificantly different than the baseline condition ($p < 0.05$). Table IV presents the statistical
 460 difference (p -values) between this combination and the baseline condition, and shows that
 461 though the users rating was the same for the index scale, they significantly preferred the
 462 baseline condition over monolog with song combination for frequent use and preference.

TABLE IV. *Post hoc* Tukey HSD Analysis: (p - values) comparing mean ratings of music-based concurrent combinations combinations scales with the relevant baseline scales: (*Signif.codes* : $\leq 0.001 = ***, \leq 0.01 = **, \leq 0.05 = *$)

Combination	Ind.	Men.	Phy.	Tem.	Eff	Fru.	Per.	Fre.	Lik.
Mon. w. Son.	0.89	**	0.28	0.07	**	*	**	***	***
Mon. w. Mus.	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99
Int. w Son.	0.99	**	0.20	**	**	*	***	***	***
Int. w Mus.	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
Com. w Son.	1.00	0.28	0.90	*	*	*	***	**	***
Com w Mus.	1.00	0.44	0.91	0.22	0.54	0.82	0.08	0.08	*
New. w. Son.	1.00	0.99	0.80	0.97	0.82	0.84	0.38	0.12	0.12
New. w. Mus	1.00	1.00	1.00	1.00	1.00	0.98	0.93	0.39	0.19
Son w. Mus.	1.00	1.00	0.99	0.98	1.00	0.96	0.77	0.25	0.12

463 In two more combinations involving song, (that is, interview with song and commentary
464 with song), users' responses were similar to the monolog with song combination. Statistical
465 analysis, mentioned in Table IV, shows that though the users rating was the same for index
466 scale, they significantly preferred the baseline condition over these combinations.

467 In the news headline with song design, intermittent news headlines were combined with
468 song. The results showed that the mean index score for listening task was 55.60. The

469 two-way ANOVA test comparing this combination with the baseline condition showed no
470 significant difference in the users' responses concerning presentation type (news headlines
471 with song — baseline), $F(1, 702) = 0.166, p < 0.684$. In *Post hoc* Tukey HSD test performing
472 the comparison between this combination and the baseline condition, no significant difference
473 ($p > 0.05$) appeared in all the rating scales. The statistical analysis showed that the user
474 experience in this combination was similar to the baseline condition.

475 In the monolog with music combination, results show that users have a greater interest
476 in this combination than the combinations previously discussed. In this combination, the
477 mean index score for the listening task was 53.38. The two-way ANOVA test comparing
478 this combination with the baseline condition shows neither a significant difference in pre-
479 sentation type (monolog with song — baseline), $F(1, 702) = 0.169, p < 0.681$, nor in the
480 interaction between the presentation and rating scales, $F(8, 702) = 1.177, p < 0.31$. Since no
481 significant impact appeared in the interaction between the presentation and rating scales, in
482 the extended analysis using *Post hoc* Tukey HSD test performing the comparison between
483 this combination and the baseline condition, no significant difference ($p = 1$) appeared in
484 all the rating scales. The statistical analysis showed that the user experience was similar to
485 the baseline condition.

486 In the other four combinations involving music (instrumental — non-vocal), that is inter-
487 view with music, commentary with music, news headlines with music and song with music,
488 similar statistical results appeared as seen in monolog with music combination discussed
489 above, see Table IV. There was one exception, as a significant difference appeared in like
490 scale in commentary with music combination. This statistical analysis shows that the user

491 experience was similar to the baseline condition in each concurrent combination involving
492 music.

493 C. Information Streams Impact in Concurrent Communication

494 Besides discussing each combination and comparing them with the baseline condition,
495 we also carried out an overall comparison regarding each information type to determine its
496 impacts when presented concurrently with rest of the information types. In other words, we
497 determined the viability of the information types to be presented concurrently with other
498 information streams.

499 In the following subsections, we discuss each of the information streams individually and
500 calculate the mean index score and mean rating for the subjective scales (as was completed in
501 the combination analysis). For each information type, we also compare the results with the
502 baseline condition statistically, following the pattern mentioned in the combination types
503 analysis. Before discussing each information type, we first present Figure 5 to show the
504 results for each information type, compared with the baseline condition depicted with a
505 continuous black line. The statistical comparison of each combination with the baseline
506 condition using *Post hoc* Tukey HSD test is mentioned in Table V.

507 In this analysis, we start with the monolog, and follow the same pattern adopted previ-
508 ously, calculating the overall mean values for index score based on the subjective subscales.
509 The ratings for the other two scales that include frequent and like are also mentioned. The
510 results showed that the index score for listening to a concurrent combination that had a
511 stream type of monolog was 60.96.

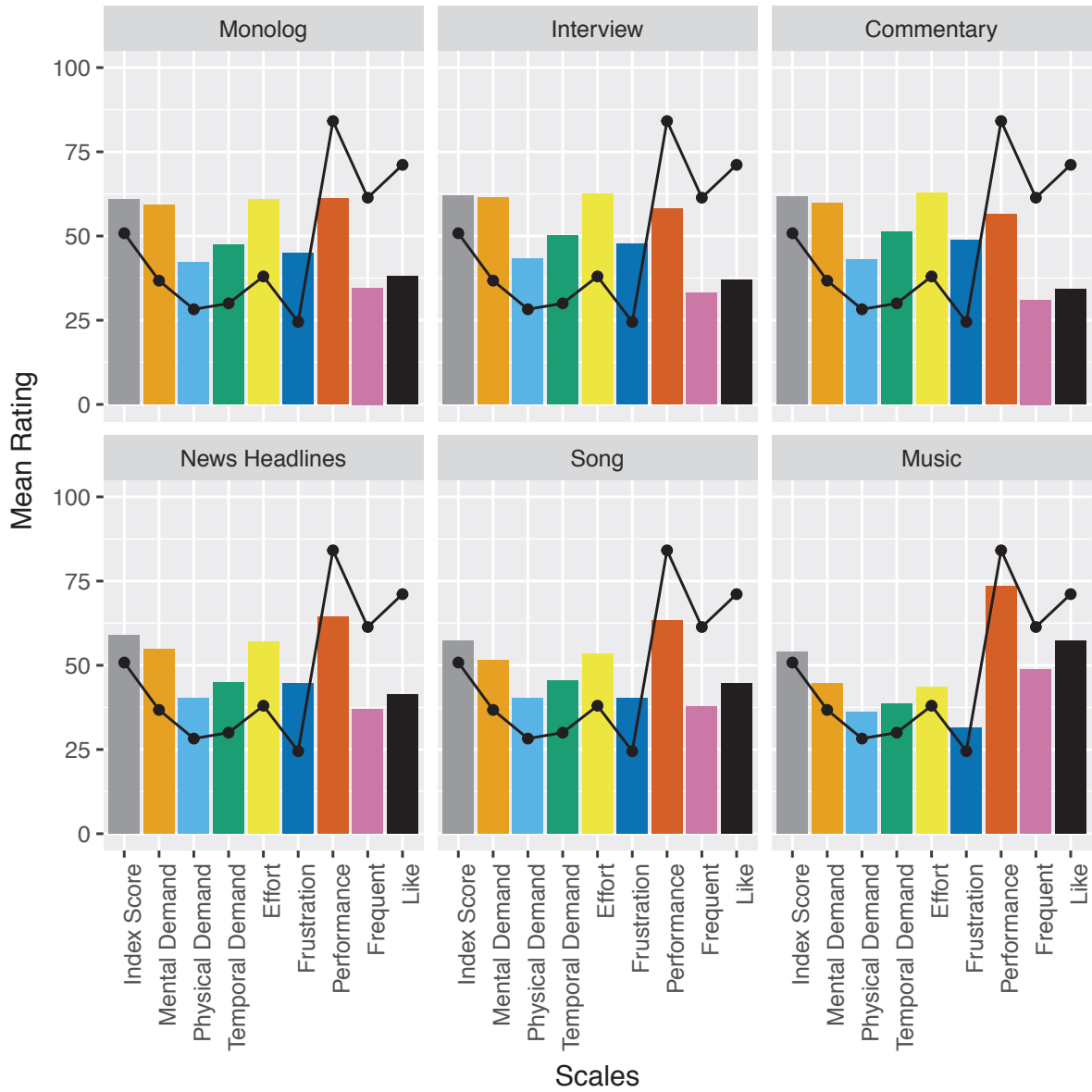


FIG. 5. Users' experience regarding each information type when presented with rest of the information types, and also compared with the baseline condition shown with a continuous line

512 In this analysis, we also performed the two way ANOVA on the results and then used *Post*
 513 *hoc* Tukey HSD test for extended analysis to determine the significant difference between the
 514 scales of information type being discussed and the baseline condition. We compared each

TABLE V. *Post hoc* Tukey HSD Analysis: (*p-values*) comparing mean ratings of each stream type with the relevant baseline scales: (*Signif.codes* : $\leq 0.001 = ***, \leq 0.01 = **, \leq 0.05 = *$)

Stream Type	Ind.	Men.	Phy.	Tem.	Eff	Frus.	Per.	Fre.	Lik.
Monolog	0.1	***	**	***	***	***	***	***	***
Interview	*	***	***	***	***	***	***	***	***
Commentary	*	***	***	***	***	***	***	***	***
News Headlines	0.46	***	*	***	***	***	***	***	***
Song	0.86	***	*	***	***	***	***	***	***
Music	1.00	0.56	0.55	0.38	0.95	0.76	0.08	*	**

515 information type with the baseline condition because in the baseline condition a randomly
516 picked stream of any type from 6 information streams used in this study was presented to the
517 users sequentially. Therefore, we compared an information type when presented concurrently
518 with the information types (baseline) that were presented sequentially.

519 For this information type, the two-way ANOVA showed that the stream type (monolog
520 — baseline) had a significant impact on users' response $F(1, 2502) = 6.643, p < 0.01$. Also,
521 the interaction between the presentation and rating scales, $F(8, 2502) = 55.155, p < 0.01$,
522 had a significant impact on the users' response. Moreover, the extended analysis using
523 *Post hoc* Tukey HSD test showed no significant difference ($p > 0.05$) regarding mean index
524 score. Table V shows the statistical difference (*p-values*) between this stream type and

525 the baseline condition. The results for music, news headlines, and song, as shown in Table
 526 V, appeared similar as seen in monolog type of stream. In all these types of streams, the
 527 index score difference was non-significant compared to the baseline condition, but users still
 528 preferred the baseline condition more than the concurrent types of information streams.

529 For the interview type, the same analysis pattern that was adopted for monolog type
 530 was followed. In this information type, the overall mean index score was 61.93. The two-
 531 way ANOVA on this information type showed that the stream type (interview — baseline)
 532 had a significant impact on the users' response $F(1, 2502) = 10.109, p < 0.001$. Also, the
 533 interaction between the presentation and rating scales, $F(8, 2502) = 62.502, p < 0.01$, had a
 534 significant impact on the users' response. The extended analysis using *Post hoc* Tukey HSD
 535 test showed significant differences ($p < 0.05$) in all the scales including mean index score
 536 ($p < 0.05$). This shows that the user experience was the least favoured in this information
 537 type when compared to the rest of the information types when presented concurrently.

538 In the commentary type of information stream, as shown in Table V, the results appeared
 539 similar to the interview type of information stream. Therefore, the user experience was the
 540 least favoured in this information type when compared to the rest of the information types
 541 when presented concurrently.

542 D. Impact of Presentation in Left — Right Ears

543 In this study, the first type of information, that is monolog streams, was always played
 544 in the left ear for all relevant concurrent combinations. Similarly, the last information
 545 stream type, that is music, was set to play in the right ear, always. However, the rest of

546 the information streams, in some combinations were presented in the left ear, and in some
547 combinations in the right ear.

548 In this study, the interview stream was once presented in the right ear, and four times it
549 was presented in the left ear of users. The commentary stream was presented twice in the
550 right ear, and the remaining three times it was presented in the left ear. Regarding news
551 headlines, it was presented in the left ear twice, and three times in the right ear. Finally,
552 the song was presented once in the left, and the remaining four times to the right ear of the
553 users in concurrent combinations. The composition is also indicated in Table I as mentioned
554 in the method section above.

555 This composition enabled us to further extended our analysis to see the users' response
556 with reference to presenting information in different ears, and to determine whether one
557 ear had an advantage over the other ear in terms of enhancing the user experience, or
558 not. For this, we compared users response regarding each information stream concerning its
559 concurrent presentation in different ears. The analysis revealed an interesting pattern and
560 showed that users reported lower workload index score for each of the information streams
561 presented in the left ear, and similarly, rated higher for frequent and like scales for left ear
562 presentation. Figure 6 and 7 show the pattern.

563 Following the same statistical pattern, a three-way ANOVA was performed which included
564 the interactions between the three independent variables in determining the significant im-
565 pact of all the independent variables on the user's response. The statistical results show that
566 the ear presentation was significant $F(1, 4428) = 7.718, p = 0.005$ $F(1, 7128) = 15.989, p < 0$
567 in impacting the user response. However, the three-way interaction between the infor-

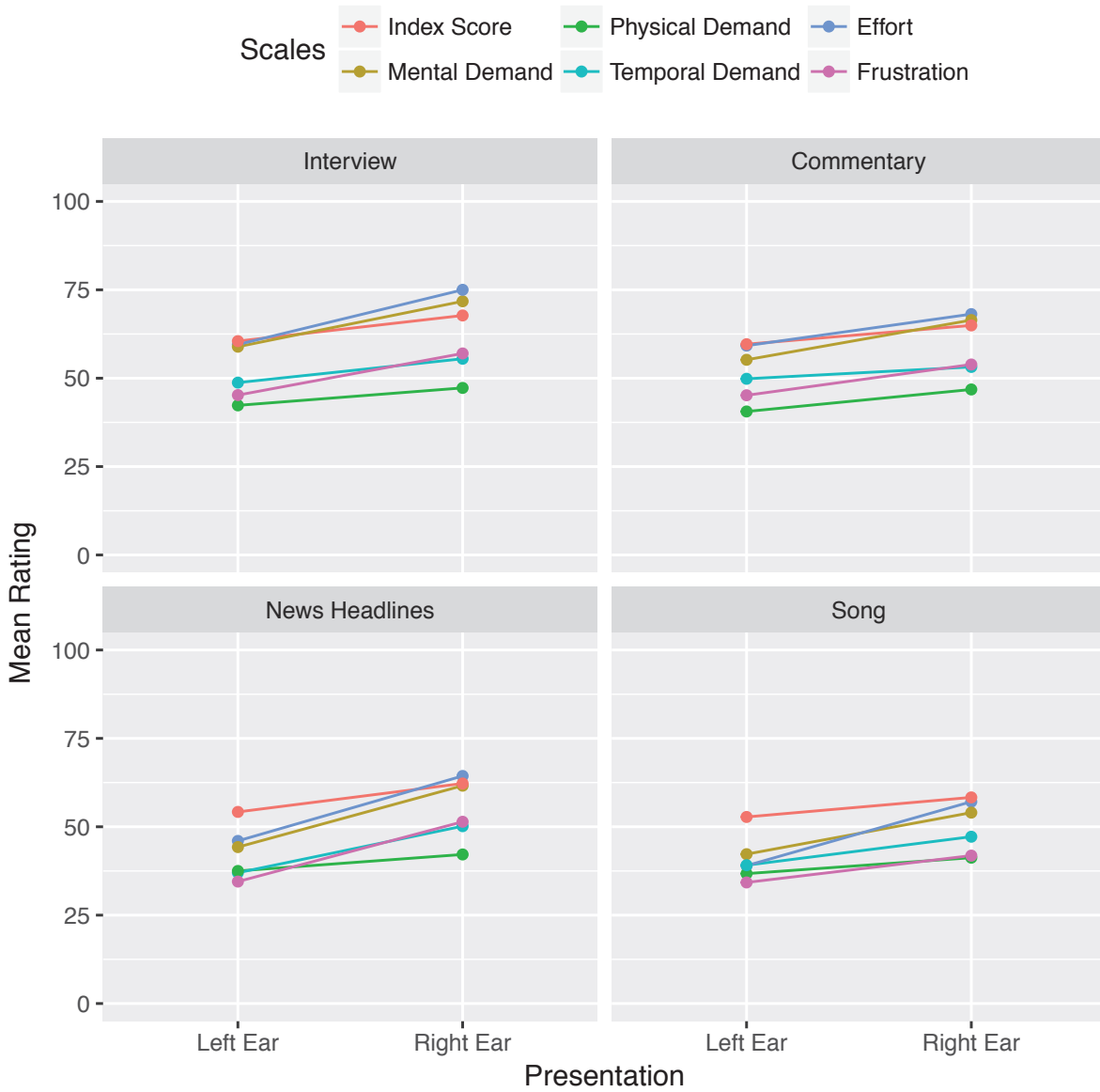


FIG. 6. Impact of Users' Experience (Stress): in each of the four information types with reference to ear presentation

568 mation type, scale type, and the ear presentation variables had a non-significant impact
 569 $F(24, 4428) = 0.392, p < 0.997$ $F(24, 7128) = 0.977, p < 0.494$ on user response, and there-
 570 fore, we did not perform the *Post hoc* Tukey HSD analysis on ANOVA results.

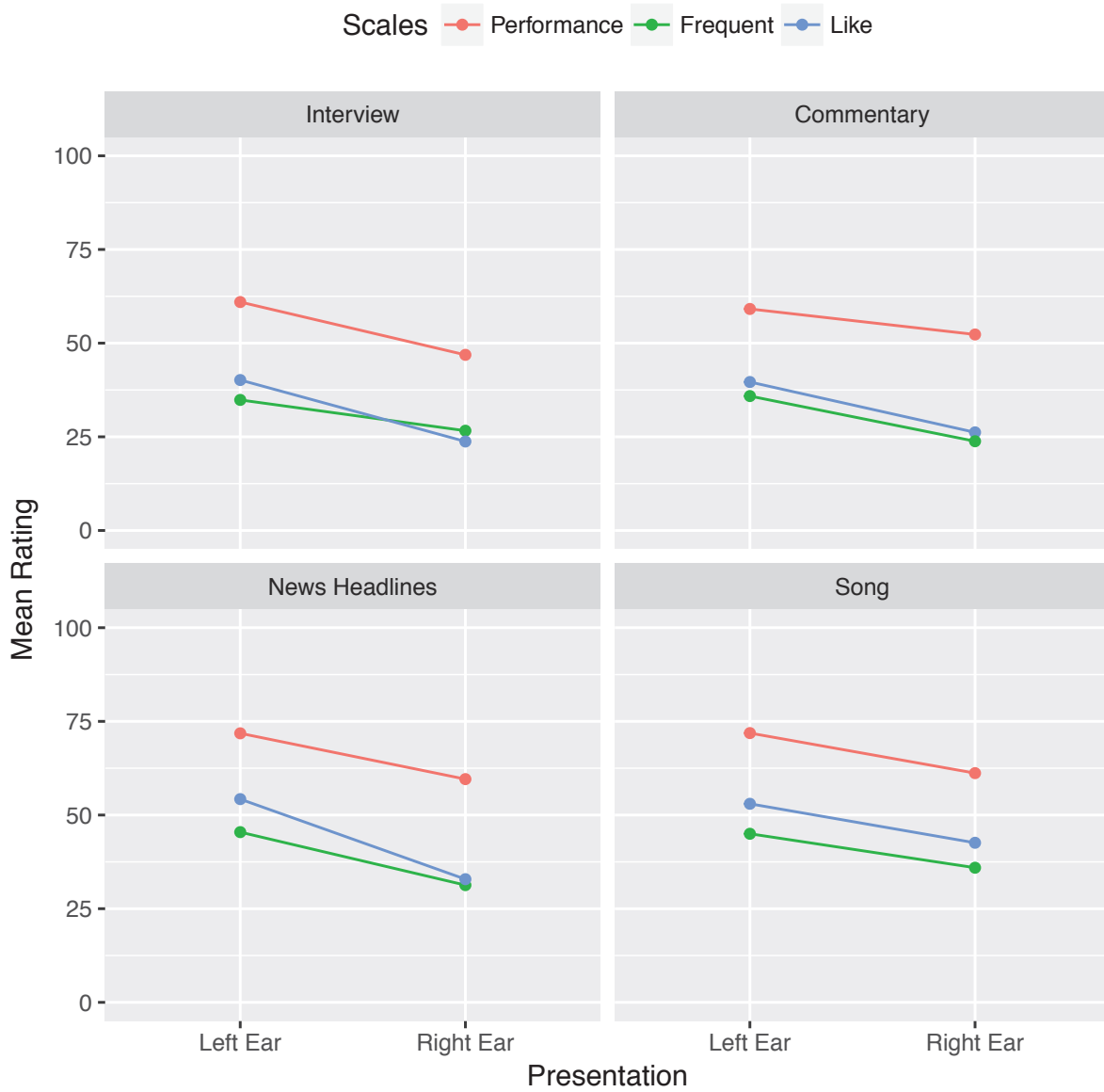


FIG. 7. **Impact of Users' Experience (Acceptance):** in each of the four information types with reference to ear presentation

571 VI. DISCUSSION

572 In our analysis, we calculated the perceived workload index score for each of the combi-
 573 nations and the baseline condition. The study showed that the perceived workload index

574 was the lowest in the baseline condition. Though there is an ascending order of concurrent
575 combinations, as shown in Figure 8, the statistical tests, as discussed in section V, showed no
576 significant difference ($p > 0.05$) between the baseline condition and each of the concurrent
577 combinations. Therefore, we conclude that the perceived workload index score in concurrent
578 and baseline conditions is not significantly different.

579 However, contrary to the results found for perceived workload index, user responses
580 in preference and frequently using different combinations were significantly different when
581 concurrent conditions were compared to the baseline condition. As shown in Figure 9, for
582 preference and frequent use, users again rated the baseline condition the highest, followed by
583 the following order of the streams with reference to frequent use. In the statistical analysis
584 for many of the streams, users' ratings regarding frequent and like scales was significantly
585 less than the baseline condition. The analysis shows that though the perceived workload
586 index remains similar in baseline condition and concurrent combinations, comparing to some
587 of the combinations, for example, interview with commentary, monolog with commentary,
588 monolog with interview, users significantly ($p < 0.05$) preferred baseline condition in terms
589 of preference and their likely frequent use.

590 The illustration in Figure 9, shows a relationship between frequent and preference (like)
591 scales and looks directly proportional to each other. The analysis also shows an inverse
592 relationship between perceived workload index score and frequent & like scales. The in-
593 versely proportional relationship between the index score and the frequent and like scales
594 is illustrated in Figure 10. This relationship shows that an increase in the perceived work-

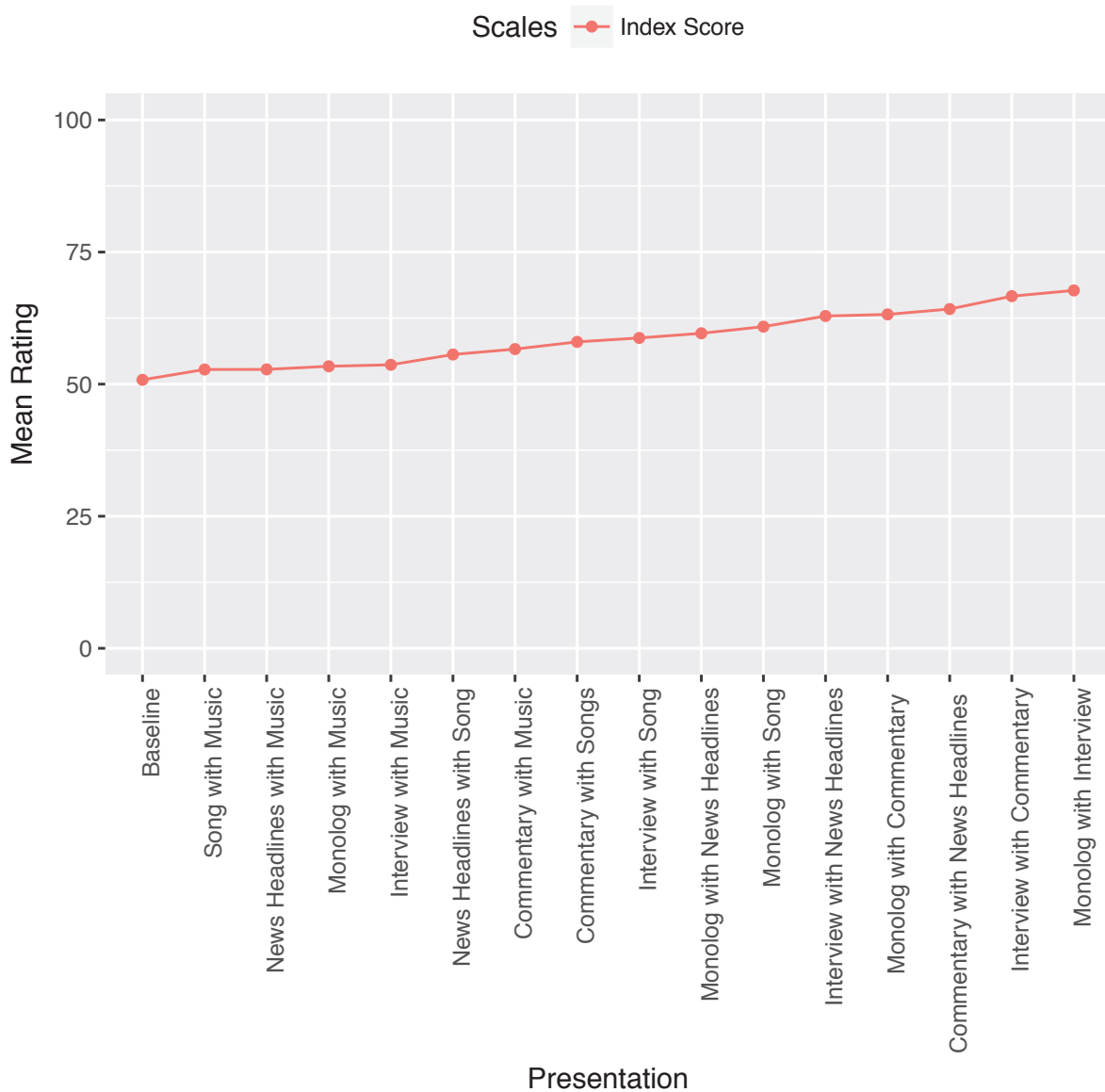


FIG. 8. Perceived Workload Index Score: an order of combinations with reference to their listening task index score reported by the users

595 load index for listening task means the relevant concurrent combination would less likely be
 596 preferred for frequent use by the users.

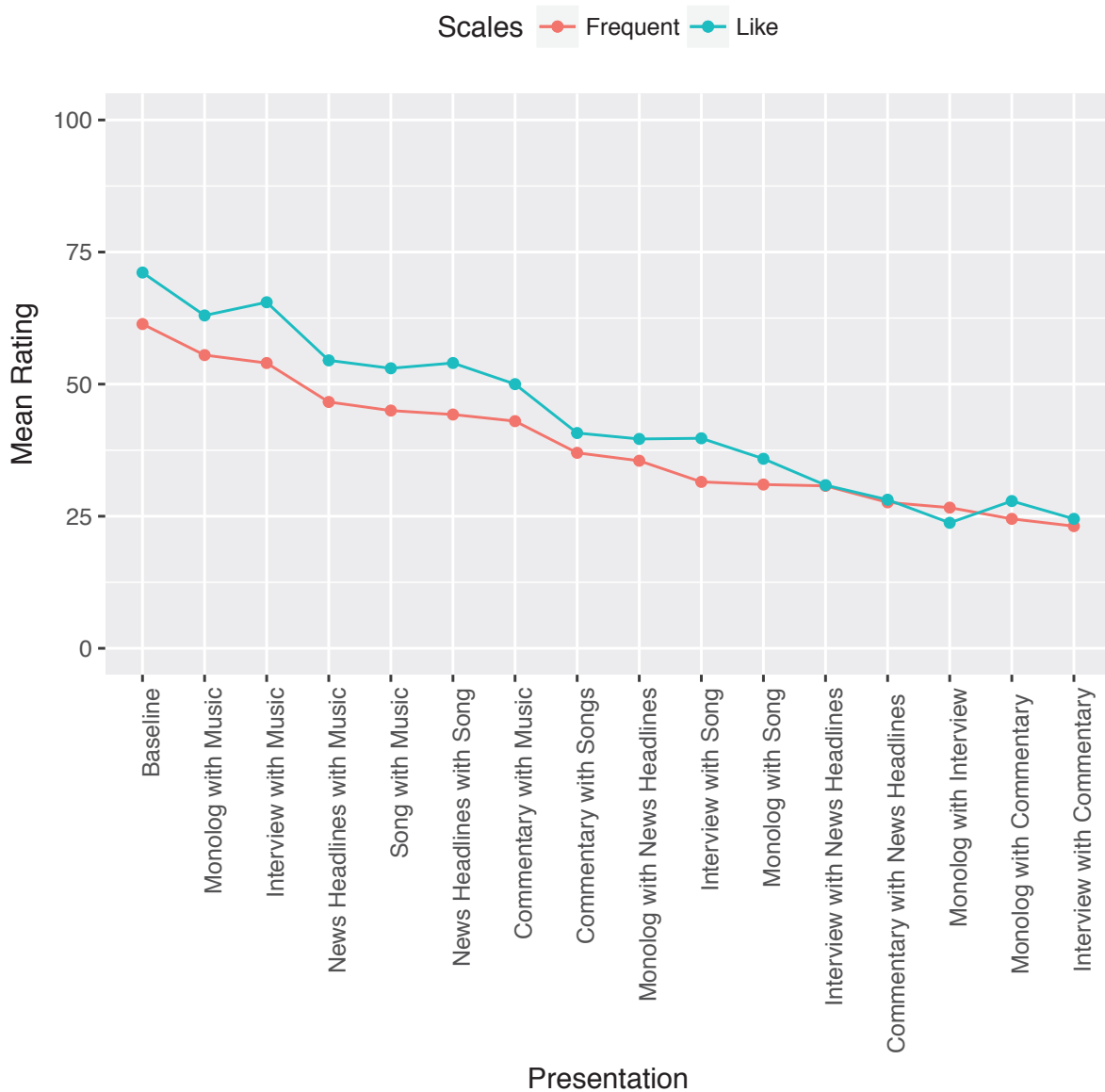


FIG. 9. Ratings for Frequent and Like (Preference) Scales: an order of combinations with reference to users' ratings regarding frequent and preference

597 The pattern in Figure 9 also suggests that the perceived workload for a listening task
 598 in concurrent combination is dependent on the type of information as well as the amount
 599 of information presented to the users. From the order of the combinations appearing in

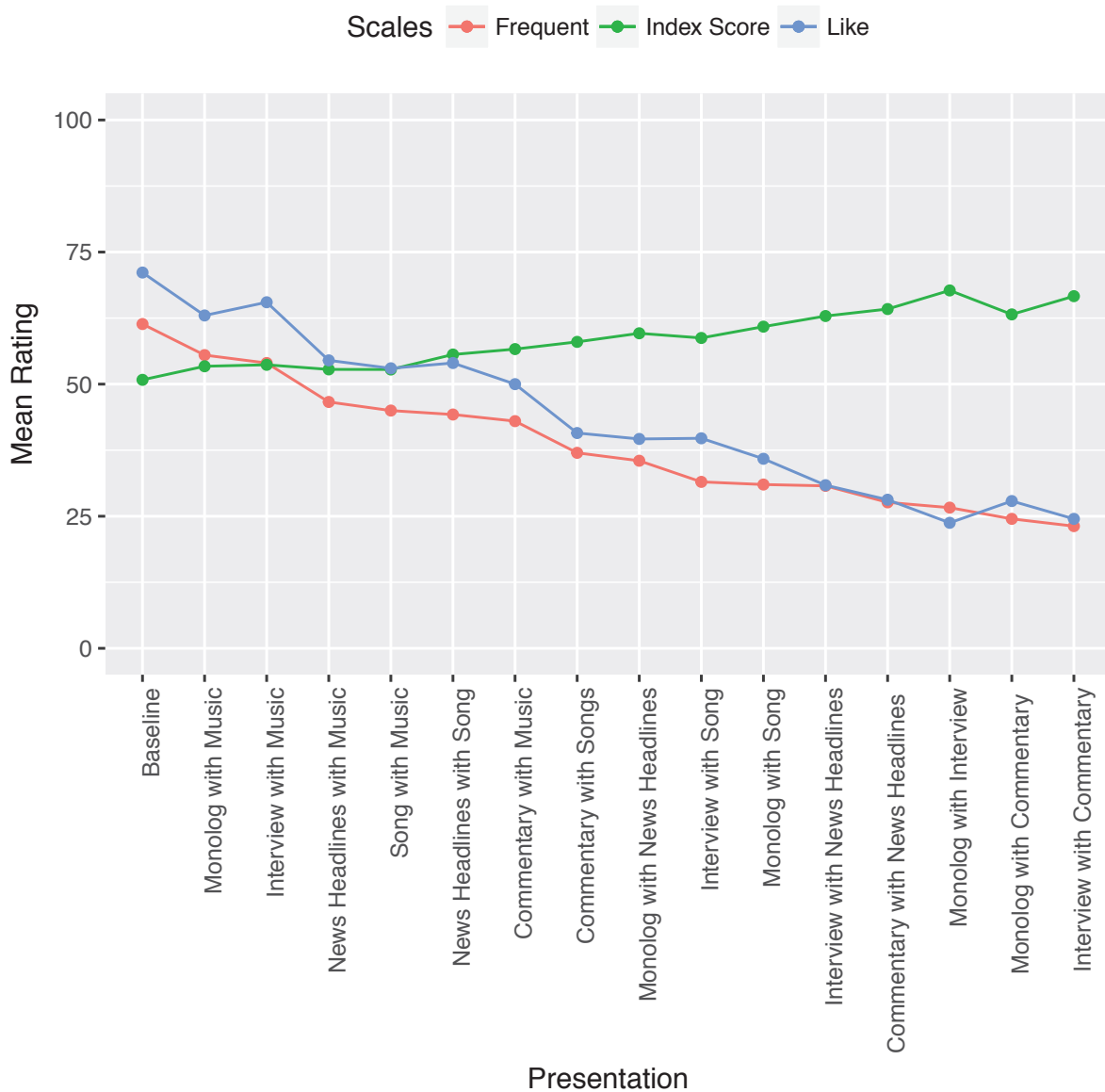


FIG. 10. **Order of Combinations:** with reference to their listening task index score, ratings for frequent and preference reported by the users

600 Figure 9, the combinations created with music were preferred the most within concurrent
 601 combinations, followed by song-based concurrent combinations. This shows, as the music
 602 and songs usually do not require focused attention to process the information stream, there

603 is apparently less cognitive load as users rated them high for frequent use. Similarly, in
604 news headlines, the controlled and limited amount of information was being provided in-
605 termittently to the users in chunks, therefore, users selected it the third highest choice to
606 hear them in all concurrent combinations. Similarly, the concurrent combinations created
607 with monolog, interview, and commentary were continuously delivering a high amount of
608 voice-based information, therefore, were rated low for frequent use by the users. This pat-
609 tern shows that the high amount of information delivery requiring greater attention and
610 cognitive processing from the users to comprehend information makes it less acceptable for
611 the users.

612 The extended analysis discussing the viability of information type in concurrent combina-
613 tion also validates the above observation that the perceived workload for a listening task in
614 concurrent combination is dependent on the type of information as well as the amount of in-
615 formation presented to the users. The order shown in Figure 11 validates the same as users
616 rated music the highest regarding frequent use when listened in concurrent combination,
617 followed by song as the second.

618 From the information type providing speech-based information (non-music/song), users
619 rated news headlines the highest for frequent use when listening to concurrent combinations.
620 Rating news headlines the highest support our previous studies that show the intermittent
621 design with the spatial difference in sources is the best form to communicate multiple in-
622 formation concurrently. In our previous study, in the intermittent form of concurrent com-
623 munication, users comprehended the content equal to the amount that they comprehended
624 in the baseline sequential presentation. In this study, users reported that intermittent form

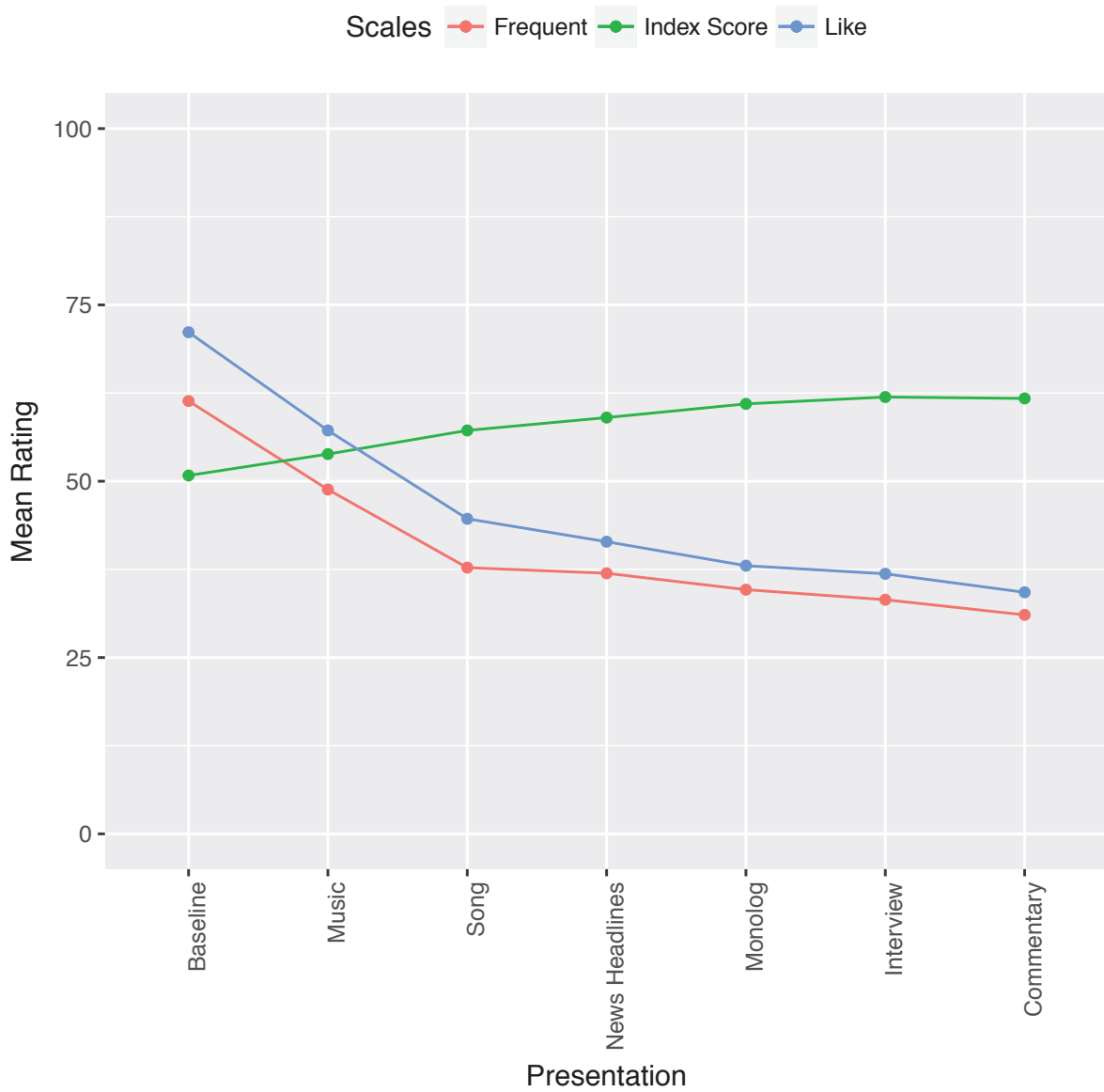


FIG. 11. Order of Information Types with reference to their potential for being a part of concurrent communication

625 of communication creates the least perceived workload index in speech-based information
 626 communication.

627 From the speech-based information types, after the news headlines, users rated monolog
628 the second highest information stream regarding frequent use. In monolog, one speaker
629 presented information, whereas in interview (dialog) two speakers were involved in presenting
630 information. Comparing these two types of information streams, users rated monolog higher
631 than the interview. This shows that the number of talkers in concurrent streams also affects
632 users perceived workload index. A stream with one talker is rated higher in terms of frequent
633 use than the stream involving two talkers.

634 Users rated commentary the least. There could be many factors, such as there being
635 background noise generated from the spectators in the commentary stream or the speaking
636 speed of the commentator being fast in order to keep up with the pace of the game. The
637 higher speaking rate/pace might also have created a difference in monolog and dialog as the
638 speaking rate in monolog was less than the dialog.

639 As discussed in the ear impact section, the detailed analysis showed that in concurrent
640 communication, users preferred an information type more when presented in the left ear as
641 compared to the right ear presentation. In all the information types, the analysis showed
642 that users reported lower workload index score for each of the information stream when it
643 was presented in the left ear, and rated higher for frequent and preference scales compared
644 to right ear presentation, and vice-versa.

645 VII. LIMITATIONS AND FUTURE WORK

646 Though this study tried to compare different combination streams comprehensively, the
647 impact on user experience by the content played and user's interest in the topics of the
648 information streams cannot be fully excluded.

649 In the analysis of the combination types, it appeared that the monolog presented with
650 the music achieved almost the same user experience as users enjoyed in the baseline condi-
651 tion. The user experience almost remained the same in baseline condition and combinations
652 with music. This shows that presenting music with an information stream does not create a
653 significant difference in user experience as compared to the baseline condition. Some users
654 rated combinations with music higher than the baseline condition in terms of frequent use
655 and preference. It sets another direction for investigation to test the impact of music pre-
656 sented in one ear while comprehending content from other voice-based streams presented
657 in the other ear. Many people do tasks with music playing in the background. Dichotic
658 listening could be tested where a speech-based information stream is provided in one ear
659 and the music stream in another ear. A study based on the same design pattern that was
660 adopted in Study ([Fazal *et al.*, 2018a](#)) could be used to compare the content comprehension
661 in such concurrent communication with the baseline sequential communication.

662 ACKNOWLEDGMENTS

663 This research was supported by the School of Computer Science, Faculty of Engineering
664 and IT, University of Technology Sydney, Australia.

675 **APPENDIX A: REFERENCES**

676

677 Audacity. “Audacity” <https://www.audacityteam.org/>, [Online; accessed 23-Dec-2018].

678 Bee, M. A., and Micheyl, C. (2008). “The cocktail party problem: what is it? how can it be
679 solved? and why should animal behaviorists study it?,” *Journal of comparative psychology*
680 **122**(3), 235.

681 Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound* (MIT
682 press).

683 Brock, D., McClimens, B., Trafton, J. G., McCurry, M., and Perzanowski, D. (2008).
684 “Evaluating listeners’ attention to and comprehension of spatialized concurrent and serial
685 talkers at normal and a synthetically faster rate of speech,” in *Proceedings of the 14th*
686 *International Conference on Auditory Display (ICAD)*, Georgia Institute of Technology,
687 pp. 1–8.

688 Brock, D., Wasylyshyn, C., McClimens, B., and Perzanowski, D. (2011). “Facilitating
689 the watchstander’s voice communications task in future navy operations,” in *MILCOM*
690 *2011 Military Communications Conference*, pp. 2222–2226, doi: [10.1109/MILCOM.2011.](https://doi.org/10.1109/MILCOM.2011.6127692)
691 [6127692](https://doi.org/10.1109/MILCOM.2011.6127692).

692 Cherry, E. C., and Taylor, W. K. (1954). “Some Further Experiments upon the Recognition
693 of Speech, with One and with Two Ears,” *The Journal of the Acoustical Society of America*
694 **26**(4), 554–559, doi: [10.1121/1.1907373](https://doi.org/10.1121/1.1907373).

- 695 Copenhaver, M. D., and Holland, B. S. (1988). “Multiple comparisons of simple effects in
696 the two-way analysis of variance with fixed effects,” *Journal of Statistical Computation*
697 *and Simulation* 1–15.
- 698 Dix, A. (2003). *Human-computer interaction* (Pearson), p. 834.
- 699 Fazal, M. A. u., Ferguson, S., and Johnston, A. (2018a). “Investigating concurrent speech-
700 based designs for information communication,” in *Proceedings of the Audio Mostly 2018*
701 *on Sound in Immersion and Emotion*, AM’18, ACM, New York, NY, USA, pp. 4:1–4:8,
702 <http://doi.acm.org/10.1145/3243274.3243284>, doi: 10.1145/3243274.3243284.
- 703 Fazal, M. A. u., Ferguson, S., and Johnston, A. (2019). “Evaluation of Information Com-
704 prehension in Speech-based Designs for Concurrent Audio Streams,” *ACM Transactions*
705 *on Multimedia Computing, Communications, and Applications (TOMM)* (-), 1–18 sub-
706 mitted.
- 707 Fazal, M. A. u., Ferguson, S., Karim, M. S., and Johnston, A. (2018b). “Concurrent Voice-
708 Based Multiple Information Communication: A Study Report of Profile-Based Users’
709 Interaction,” in *145th Convention of the Audio Engineering Society*, Audio Engineering
710 Society.
- 711 Fazal, M. A. u., and Shuaib Karim, M. (2017). “Multiple information communication in
712 voice-based interaction,” in *Advances in Intelligent Systems and Computing* (Springer),
713 pp. 101–111, doi: 10.1007/978-3-319-43982-2_9.
- 714 Feltham, F., and Loke, L. (2017). “Felt sense through auditory display: A design case study
715 into sound for somatic awareness while walking,” in *Proceedings of the 2017 ACM SIGCHI*
716 *Conference on Creativity and Cognition*, ACM, pp. 287–298.

- 717 Guerreiro, J. (2013). “Using simultaneous audio sources to speed-up blind people’s web
718 scanning,” in *Proceedings of the 10th International Cross-Disciplinary Conference on Web*
719 *Accessibility*, ACM, pp. 1–2, doi: [10.1145/2461121.2461154](https://doi.org/10.1145/2461121.2461154).
- 720 Guerreiro, J. (2016). “Enhancing blind peoples information scanning with concurrent
721 speech,” Ph.D. thesis, University of Lisbon.
- 722 Guerreiro, J., and Goncalves, D. (2016). “Scanning for digital content: How blind and
723 sighted people perceive concurrent speech,” *ACM Transactions on Accessible Computing*
724 **8**(1), doi: [10.1109/CVPR.2016.105](https://doi.org/10.1109/CVPR.2016.105).
- 725 Hart, S. G., and Stavenland, L. E. (1988). “Development of NASA-TLX (Task Load Index):
726 Results of empirical and theoretical research,” in *Human Mental Workload*, edited by
727 P. A. Hancock and N. Meshkati (Elsevier), Chap. 7, pp. 139–183, [http://ntrs.nasa.](http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000004342_1999205624.pdf)
728 [gov/archive/nasa/casi.ntrs.nasa.gov/20000004342_1999205624.pdf](http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20000004342_1999205624.pdf).
- 729 Hinde, A. F. (2016). “Concurrency in auditory displays for connected television,” Ph.D.
730 thesis, University of York.
- 731 Ikei, Y., Yamazaki, H., Hirota, K., and Hirose, M. (2006). “vCocktail: multiplexed-voice
732 menu presentation method for wearable computers,” in *Virtual Reality Conference*, IEEE,
733 pp. 183–190, doi: [10.1109/VR.2006.141](https://doi.org/10.1109/VR.2006.141).
- 734 Miller, R. G. (198). *Simultaneous Statistical Inference* (Springer).
- 735 Mullins, A. T. (1996). “Audiostreamer: Leveraging The Cocktail Party Effect for Efficient
736 Listening,” Ph.D. thesis, Massachusetts Institute of Technology.
- 737 NASA (2018a). “NASA Task Load Index Sheet” [https://humansystems.arc.nasa.gov/](https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf)
738 [groups/TLX/downloads/TLXScale.pdf](https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf), [Online; accessed 22-Dec-2018].

- 739 NASA (2018b). “NASA-TLX” <https://humansystems.arc.nasa.gov/groups/TLX/>, [Online;
740 accessed 22-Dec-2018].
- 741 Neil, T. (2009). *Designing web interfaces: Principles and patterns for rich interactions*
742 (O’Reilly Media, Inc.).
- 743 Parente, P. (2008). “Cliques: Perceptually based, task oriented auditory display for GUI
744 applications,” Ph.D. thesis, The University of North Carolina at Chapel Hill.
- 745 Sanderson, P. (2006). “The multimodal world of medical monitoring displays,” *Applied*
746 *Ergonomics* **37**(4), 501–512, doi: [10.1016/j.apergo.2006.04.022](https://doi.org/10.1016/j.apergo.2006.04.022).
- 747 Sato, D., Zhu, S., Kobayashi, M., Takagi, H., and Asakawa, C. (2011). “Sasayaki: aug-
748 mented voice web browsing experience,” in *Proceedings of the SIGCHI Conference on*
749 *Human Factors in Computing Systems*, ACM, pp. 2769–2778.
- 750 Schmandt, C. (1998). “Audio hallway: a virtual acoustic environment for browsing,” in *Pro-*
751 *ceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*,
752 ACM, pp. 163–170, doi: [10.1145/288392.288597](https://doi.org/10.1145/288392.288597).
- 753 Schmandt, C., and Mullins, A. (1995). “AudioStreamer: Exploiting simultaneity for listen-
754 ing,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*,
755 ACM, pp. 218–219, doi: [10.1145/223355.223533](https://doi.org/10.1145/223355.223533).
- 756 Towers, J. A. (2016). “Enabling the Effective Application of Spatial Auditory Displays in
757 Modern Flight Decks,” Ph.D. thesis, The University of Queensland.
- 758 Truschin, S., Schermann, M., Goswami, S., and Krcmar, H. (2014). “Designing interfaces for
759 multiple-goal environments,” *ACM Transactions on Computer-Human Interaction* **21**(1),
760 1–24, doi: [10.1145/2544066](https://doi.org/10.1145/2544066).

- 761 ul Fazal, M. A., Ferguson, S., and Johnston, A. (2019). “Multiple Speech-based Infor-
762 mation Communication,” [https://figshare.com/articles/Multiple_Speech-based_](https://figshare.com/articles/Multiple_Speech-based_Information_Communication/8214776)
763 [Information_Communication/8214776](https://figshare.com/articles/Multiple_Speech-based_Information_Communication/8214776), doi: [10.6084/m9.figshare.8214776.v1](https://doi.org/10.6084/m9.figshare.8214776.v1).
- 764 Vazquez-Alvarez, Y., Aylett, M. P., Brewster, S. A., Jungenfeld, R. V., and Virolainen,
765 A. (2015). “Designing interactions with multilevel auditory displays in mobile audio-
766 augmented reality,” *ACM Transactions on Computer-Human Interaction* **23**(1), 1–30.
- 767 Vazquez-Alvarez, Y., Aylett, M. P., Brewster, S. A., von Jungenfeld, R., and Virolainen,
768 A. (2014). “Multilevel auditory displays for mobile eyes-free location-based interaction,”
769 in *Proceedings of the Extended Abstracts of the 32Nd Annual ACM Conference on Human*
770 *Factors in Computing Systems*, ACM, pp. 1567–1572.
- 771 Vazquez Alvarez, Y., and Brewster, S. A. (2010). “Designing spatial audio interfaces to
772 support multiple audio streams,” in *Proceedings of the 12th International Conference on*
773 *Human Computer Interaction with Mobile Devices and Services*, ACM, pp. 253–256, doi:
774 [10.1145/1851600.1851642](https://doi.org/10.1145/1851600.1851642).
- 775 Walter, S. R., Raban, M. Z., Dunsmuir, W. T., Douglas, H. E., and Westbrook, J. I. (2017).
776 “Emergency doctors’ strategies to manage competing workload demands in an interruptive
777 environment: An observational workflow time study,” *Applied Ergonomics* **58**, 454–460,
778 doi: [10.1016/j.apergo.2016.07.020](https://doi.org/10.1016/j.apergo.2016.07.020).
- 779 Ware, C. (2012). *Information visualization: perception for design* (Elsevier).
- 780 Werner, S., Hauck, C., Roome, N., Hoover, C., and Choates, D. (2015). “Can VoiceScapes
781 assist in menu navigation?,” in *Proceedings of the Human Factors and Ergonomics Society*,
782 Vol. 2015, pp. 1095–1099, doi: [10.1177/1541931215591157](https://doi.org/10.1177/1541931215591157).

- 783 Xia, J., Nooraei, N., Kalluri, S., and Edwards, B. (2015). “Spatial release of cognitive load
784 measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners,” *The*
785 *Journal of the Acoustical Society of America* **137**(4), 1888–1898, doi: [10.1121/1.4916599](https://doi.org/10.1121/1.4916599).
- 786 Yandell, B. S. (1997). *Practical Data Analysis for Designed Experiments* (Chapman Hall).

Appendix O

Publication 9

A. Hussain, M. A. u. Fazal, and M. S. Karim, "Intra-domain user model for content adaptation," in *Smart Innovation, Systems and Technologies*. Springer, 2015, pp. 285–295

Intra-domain User Model for Content Adaptation

Anwar Hussain¹, M. Abu Ul Fazal², and M. Shuaib Karim³

¹ Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan
anwarh45@gmail.com

² Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan
fazalsidhu@yahoo.com

³ Department of Computer Sciences, Quaid-i-Azam University, Islamabad, Pakistan
skarim@qau.edu.pk

Abstract. In learning environment, personalization of contents according to the requirement of an individual student is the most important feature of adaptive educational systems. This process becomes more effective if the system knows the way through which a student learns best. Learning styles are non-stationary and are varied for academic disciplines. Our proposed model considers its non-deterministic nature, effect of the subject domain, and non-stationary aspects during the learning process. Presented approach is novel, simple but more flexible that dynamically and accurately adjusts students learning style variations in a discipline-wise manner. For the evaluation of our proposed model, Visual/Verbal dimension of Felder and Silvermen learning style model is utilized for personalization of Computer Science undergraduate subjects in our experimental prototype. Results show that personalization of contents in a discipline-wise manner is more effective during the learning process of a student.

Keywords: E-Learning, User Modeling, Personalization, Learning Styles, Adaptive Educational Systems, Content Adaptation

1 Introduction

One of the important issues nowadays is to provide contents according to the user requirements. In addition, these contents must be presented according to their preferences. For better adaptation, a system should be aware of a different aspects of a user, e.g. requirements, needs, goals, preferences, learning styles. In educational environment, these systems provide educational contents to the students and are known as adaptive educational hypermedia systems [1, 2]. Researchers have utilized different models of human cognition [2, 3] in this field. One of the properties of learning style is its non-stationary nature that can change time to time. For this purpose, researchers are trying to propose models that are capable of updating in terms of learning styles [4, 5]. Keeping user model updated is a complex task because learning is a continuous process and human

cognition is not easily quantifiable. To our knowledge, discipline-wise learning style variations are not taken into account for personalization. The following scenario highlights the importance of discipline-wise personalization.

A student uses an adaptive system for his learning that provides course content in the form of text and videos. For example, a student wants to study *Database* course and he prefers written text. The adaptive system will develop a model and will prefer him written text. After some time, the student wants to study *Programming* course and prefers video lectures. Adaptive system will update his model and will prefer video lectures instead of written text. Consider the case if he studies *Database* course again. Adaptive system will again reverse its mechanism. The problem will arise again if he switches once more to *Programming* course again. One of the solutions to this problem is, there should be a model that could handle learning style variations in a discipline-wise manner.

In this study, we have tried to exploit the effect of the subject domain of learning styles too. Our model dynamically and continuously adjusts student learning styles preferences for multiple academic disciplines.

2 Background and Related work

During the evolution of Computer Science field, interactions between humans and computers have also evolved to increase the usability of computer systems. The first requirement for this is to understand the human. To understand a human, theories of psychology have a vital role that provide information about human's cognitive abilities. These include how a human learns, thinks and processes information. Researchers have come to the conclusion that for better adaptation, along with the physical capabilities, it is more important to consider cognitive abilities of human's [6]. A number of models of human cognition have been developed and used in the field of adaptive education for learning purposes [2, 3, 7]. For example Index of learning style, Field dependent/Field-independent, Verbal-Imager/Holistic-Analytical and Kolb's learning style inventory, which are used most frequently.

In the field of adaptive educational hypermedia systems, learning style models have an important role [2, 3]. The main aim of these studies is to enhance students' learning in a more effective way. These studies give enough information about the learning styles used in the past, modeling approaches, key variables considered and the systems developed for adaptation. Recent work in this field, most of these approaches are in the context of a single domain [4, 5, 8].

Exploiting learning styles in adaptive systems is an active area of research [4, 8]. The main focus of recent studies is to improve user performance and academic achievement. One of the important factors is that a student has variations in difficulty for different academic disciplines [9]. For better adaptation of educational contents, it is important for a user model to have learning style variations according to the requirement of academic disciplines.

3 SWA Model

SWA (Subject-wise Adaptive) is the extension of an existing model presented by [10], for accommodating discipline-wise learning style variation. For personalization of contents and users' learning style preferences, dimensions of Index of learning style model were exploited.

3.1 User's Categories and Their Probabilities

Index of learning style model categorizes users in four dimensions, where each dimension has two further types of user preferences. These preferences are Active (A) vs Reflective (R), Sensing (S) vs Intuitive (I), Visual (Vi) vs Verbal (Ve) and Global (G) vs Sequential (Seq). For each category/dimension, a user will get one of the preferences, i.e. a user will either be Active (A) or Reflective (R) but not both. Learning style of a user is the combination of preferences for dimension a , b , c , and d as shown in Equation 1. As a result, we have total 16 number of possible learning styles.

$$LSC = \{a \in (A/R), b \in (S/I), c \in (Vi/Ve), d \in (G/Seq)\} \quad (1)$$

In learning style model, each preference in a dimension has a certain probability. The total probability of a dimension is 1 while preferences in it share part of this. If probability of *Active* preference is w then probability of *Reflective* preference becomes $1 - w$, while the total probability of a dimension remains 1 i.e. $w + (1 - w)$ as shown in Table 1. Where Pr represents probability while subscripts A, R, S, I, Seq, G, Vi and Ve represents the preferences in a dimension.

Table 1. Probabilities of Preferences Inside Dimensions

Dimension	Preference (A)	Preference (B)	Total Probability (A + B)
a	$Pr_A = w$	$Pr_R = 1 - w$	$Pr_A + Pr_R = 1$
b	$Pr_S = x$	$Pr_I = 1 - x$	$Pr_S + Pr_I = 1$
c	$Pr_{Vi} = y$	$Pr_{Ve} = 1 - y$	$Pr_{Vi} + Pr_{Ve} = 1$
d	$Pr_{Seq} = z$	$Pr_G = 1 - z$	$Pr_{Seq} + Pr_G = 1$

Selection of a preference in a dimension is based on the number of answers which favor that preference. The general formula, for calculating probability of a preference is obtained from Equation 2.

$$Pr_i = \frac{A_i}{11} \quad (2)$$

Where Pr_i is the probability of preference in i^{th} dimension. A_i is the number of favorable answers for preference A in a dimension i . Here the number 11 shows the total number of questions for a dimension. Through this way, the preferences of a user in dimensions can be found.

3.2 Updating User Model Subject-wise

The criteria for updating user preferences is the performance of a student in the tests [8, 4]. In the case of lower performance, it is assumed that learning style preferences stored in the user model are not the actual representation of user preferences. Learning style preferences are updated after each unsuccessful learning session. In the beginning, preferences' values are stored same for all disciplines but later these preferences are updated in a discipline-wise manner.

3.3 Adaptation Rules

User model is updated after learning session by a number of rules if desirable performance m is not achieved by a learner in j^{th} subject. When a student fails the test, probability of existing learning style preferences are decremented by factor $R[s_j]$ and probabilities of missing preferences are incremented by the same factor $R[s_j]$. Where value of $R[s_j]$ is determined by Equation 3. These rules are represented in Table 2. In these rules

Table 2. Rules Used in Adaptation Process

<p>Rule 1 $IF(PFM[s_j] < m) \text{ AND } (LSC[d_i][s_j] = "A") \text{ THEN}$ $SM[d_i][s_j]_A = SM[d_i][s_j]_A - R[s_j]$ $SM[d_i][s_j]_B = SM[d_i][s_j]_B + R[s_j]$</p> <p>Rule 2 $IF(PFM[s_j] < m) \text{ AND } (LSC[d_i][s_j] = "B") \text{ THEN}$ $SM[d_i][s_j]_A = SM[d_i][s_j]_A + R[s_j]$ $SM[d_i][s_j]_B = SM[d_i][s_j]_B - R[s_j]$</p>

Where m is performance threshold i.e. 60, i represents dimensions of learning style model [1 – 4], j is the subject numbered from [1 – n], performance in a subject is represented by $PFM[s_j]$ in the range [0 – 100], value of preference for model i dimension for subject j is $LSC[d_i][s_j]$, $SM[d_i][s_j]_A$ is the value of preference A stored in the dimension i for subject j , and $R[s_j]$ is the value of reinforcing for subject j , where $R[s_j]$ is in [0 – 1] and is not necessarily to be equal to $R[s_k]$ for where $k = 1 \dots n$ and $j \neq k$.

$$R[s_j] = \frac{1}{PFM[s_j] * DLS[s_j]} \quad (3)$$

$DLS[s_j]$ is the distance between preferences in a dimension. DLS can be calculated by the Equation 4. Reinforcement $R[s_j]$ is inversely proportional

$PFM[s_j]$ and $DLS[s_j]$. Inverse relationship between $R[s_j]$ and $DLS[s_j]$ is necessary because when the difference between preferences is too small then these preferences are not considered stronger on either side which become insignificant or undefined preferences. All preferences in four dimensions will be updated for appropriate subject if desirable performance is not achieved in tests.

$$DLS[s_j] = |SM[d_i][s_j]_A - SM[d_i][s_j]_B| \quad (4)$$

The values of R should be in a range for which preferences should not change very rapidly or very slowly. Through simulation, it was identified that maximum value for 0.20 and smallest 0.05 for R is more suitable [11, 10]. We have used these upper and lower limits for the value of R .

4 Experimental Prototype Application

We have developed a web-based application prototype for experimentation and evaluation of proposed model using PHP, MySQL server, JavaScript, CSS and related web-based constructs.

4.1 Dataset and Learning Style Model

We have selected videos and textual documents of two Computer Science subjects from Virtual University Website. Subjects include *Database Management System* and *Introduction to Programming*. Videos and textual documents were divided into small parts i.e. topics. As a result, for each topic we have two forms of representation i.e., visual and written. Index of learning style model is exploited where we have used its *Visual/Verbal* dimension because it is more suitable for coursework hypermedia [12]. For identifying learning style preferences, questions associated with visual and verbal dimension are selected from the Index of learning style questionnaire⁴. For learning style's dimensions and their associated questions, we can see the work of [13].

4.2 Main Components of Prototype

Our prototype application has main four components. The overall architecture of the experimental prototype is shown in Figure 1.

- **Database of educational contents:** For each topic, videos and textual documents are stored in a database.
- **User model:** It is the collection of information about students' preferences for each subject.
- **Adaptation Module:** This module updates preferences when a student fails the test at the end of a lecture, using rules as depicted in Table 2.

⁴ <http://www.engr.ncsu.edu/learningstyles/ilsweb.html>

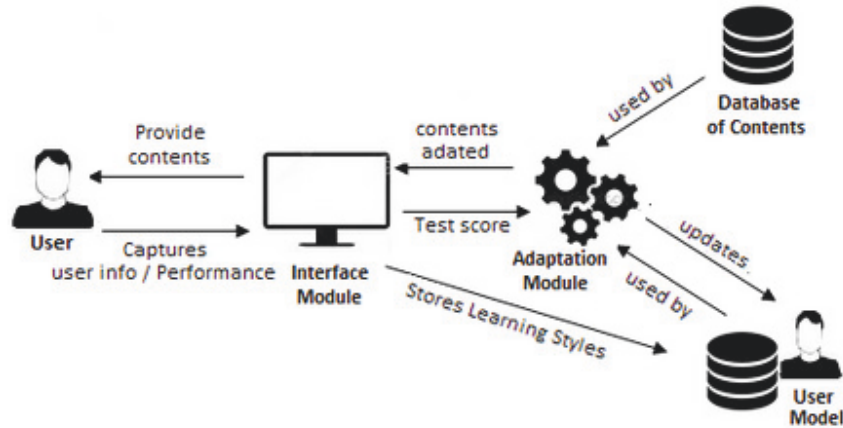


Fig. 1. Architecture of Prototype

- **Interface Module:** Interface module provides an interface which captures learning style of a student and presents contents selected by adaptation module. Another responsibility of this module is assessment of student’s performance by presenting tests after each lecture. Interface layout for visual preferences and written preferences is shown in Figure 2.

4.3 Adaptation Mechanism

- **Capturing Student Preferences:** For a new user, learning style questionnaire is given. When a student fills and submits it, preferences for the user are stored.
- **Providing Contents:** Contents are given to the student according to his preferences associated with each subject. These preferences could be different for different subjects.
- **Testing User Performance:** At the end of a lecture, the student is asked to give the test. Next lecture will be provided if the student’s score on the test is satisfactory otherwise preferences will be updated only for the subject to which the lecture belongs.
- **Updating User Preferences:** Based on the test marks, preferences associated with the current subject are updated.

5 Research Design and Evaluation

Our research design is experimental and followed in recent *learning style based user modeling approaches* [4, 5]. We are using Two-Group Pretest-Posttest Randomized Experimental Design.

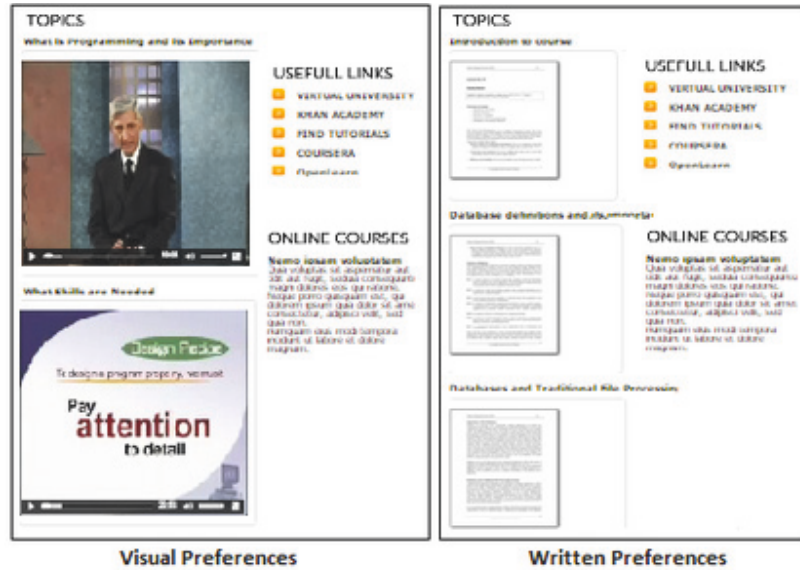


Fig. 2. Interface for Visual and Written Preferences

5.1 Experimental Setup

Evaluation of our model is based upon the comparison between our application prototype and educational website of *Virtual University*⁵ that does not have the mechanism for adaptation in discipline-wise manner. Learning style questionnaire was given to the students from *Virtual University of Pakistan* Islamabad campus. These undergraduate *Bachelor* and *Master* students were selected from *Computer Science Department*. Total 30 students were selected, 15 of them were visual while another 15 having preferences for written text. Their preferences were identified through pre-test questionnaire, i.e., learning style questionnaire.

Half of visual user were randomly assigned to the control group and other half visual users to the treatment group. The same procedure was used for the written preferences students too. We have also applied independent sample T-test for comparing means of both groups for confidence interval 95%, for the alpha value 0.05. Value of p was obtained 0.860, i.e., greater than 0.05 which means that both groups are not statistically different. Control group used website of Virtual University for learning 6 lectures, 3 per subject for the duration of one week while Treatment group used the prototype application for learning the same task. At the end of the week, post-survey questionnaire was given to both groups to record their responses to measure efficiency, effectiveness and user satisfaction about these systems. Total 13 questions were used for these variables. The detail of the questions and their purpose is shown in Table 3.

⁵ <http://www.vu.edu.pk/>

Table 3. Key Measures for Evaluation

S.No	Measure	Aspects
1	Efficiency	Identification of topics of a lecture
		Switching among topics of a lecture
2	User Satisfaction	Overall perspective
		Presentation of contents
		Scalability
		Future usage
3	Effectiveness	Increasing motivation
		Student assessment
		Usefulness
		Identifying weakness in subjects' lectures
		Help to overcome weaknesses
		Memorization of topics
4	Suggestions	Suggestions to improve the application in future

5.2 Results

Comparison between these two groups shows that our proposed model is better and effective for the students during learning. These comparisons are shown in Figure 3, Figure 4, and Figure 5 where we have plotted *above average* and *very good* responses for evaluation measures.

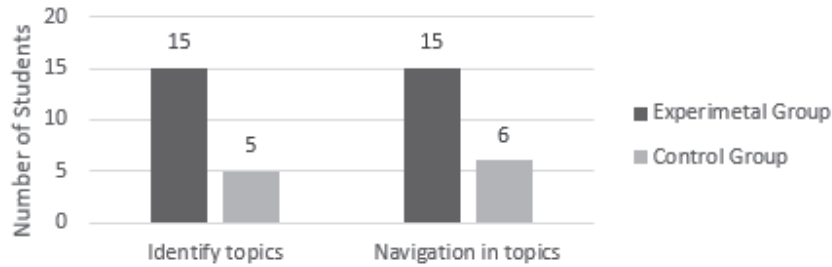


Fig. 3. Efficiency Comparison between Groups

Figure 3 shows the efficiency comparison. From Figure 3 we can see that our experimental prototype saves significant time for the tasks *identification of topics* and *navigation among topics*.

Effectiveness of the application prototype is depicted in Figure 5. We can see that most of the experimental group students agree that our system is better in *pinpointing their weakness* and system presentation help them to *work on their weaknesses*. They believe that *marking is unbiased, helpful and useful* and *increase student's motivation* during the learning process, and system presentation

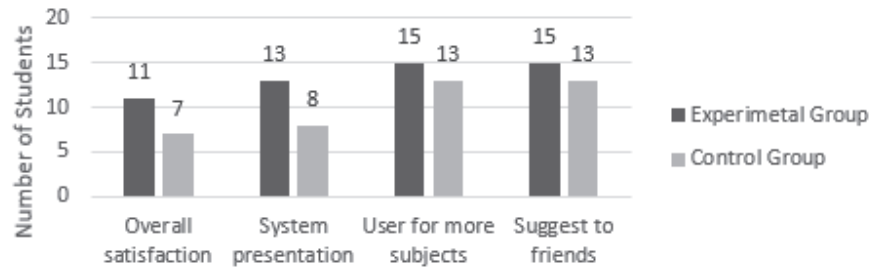


Fig. 4. User Satisfaction Comparison Between Groups

also helps them not to remember topics of a lecture if they need to study these again.

Satisfaction of user for their used system are in Figure 4. We can see from Figure 4 that most of the students are satisfied with the system's presentation, want to use it for more subjects and to suggest our application to their friends. We can see from these statistics that our proposed model for personalization of contents in term of discipline-wise learning style variations is more effective.

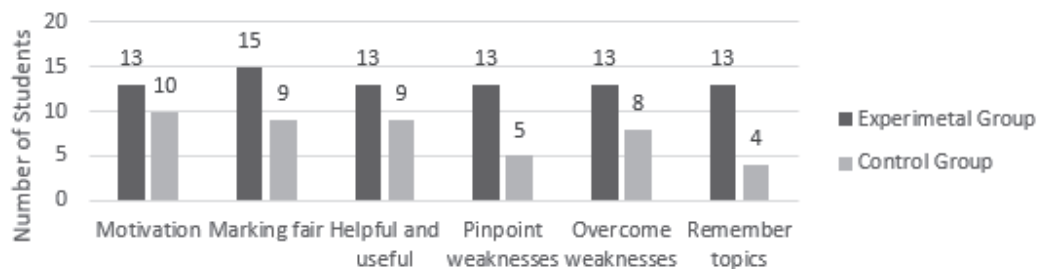


Fig. 5. Effectiveness Comparison between Groups

6 Conclusion and Future Work

Learning style based user models try to modify/adjust learning style preferences time to time to increase student performance through effective personalization. In our work, we have provided a user model that keeps discipline-wise learning style variations of a student. Our model continuously, dynamically and automatically adjusts learning style preferences for each academic discipline accordingly. Personalization becomes more effective if a user model is aware of discipline-wise requirements of the student. Results show improvements during the learning process. It is important to exploit other dimensions of learning style model

for adaptation. We are planning to evaluate the system for a larger number of participants and for a lengthy period of time. It is necessary to investigate the relationship between subject nature and learning style preferences in order to develop more mature adaptive systems for learning purposes.

References

1. C Froschl. User modeling and user profiling in adaptive e-learning systems. *Graz, Austria: Master Thesis*, 2005.
2. Evgeny Knutov, Paul De Bra, and Mykola Pechenizkiy. Ah 12 years later a comprehensive survey of adaptive hypermedia methods and techniques. *New Review of Hypermedia and Multimedia*, 15(1):5–38, 2009.
3. Catherine Mulwa, Seamus Lawless, Mary Sharp, Inmaculada Arnedillo-Sanchez, and Vincent Wade. Adaptive educational hypermedia systems in technology enhanced learning: a literature review. In *Proceedings of the 2010 ACM conference on Information technology education*, pages 73–84. ACM, 2010.
4. Hazem M El-Bakry and Ahmed A Saleh. Adaptive e-learning based on learner's styles. *Bulletin of Electrical Engineering and Informatics*, 2(4):240–251, 2013.
5. Hazem M El-Bakry, Ahmed A Saleh, Taghreed T Asfour, and Nikos Mastorakis. A new adaptive e-learning model based on learner's styles. In *Proc. of 13th WSEAS Int. Conf. on Mathematical and Computational Methods In Science and Engineering (MACMESE'11). Catania, Sicily, Italy*, pages 440–448, 2011.
6. Evelyn P Rozanski and Anne R Haake. The many facets of hci. In *Proceedings of the 4th conference on Information technology curriculum*, pages 180–185. ACM, 2003.
7. Elizabeth J Brown, Tim J Brailsford, Tony Fisher, and Adam Moore. Evaluating learning style personalization in adaptive systems: Quantitative methods and approaches. *Learning Technologies, IEEE Transactions on*, 2(1):10–22, 2009.
8. Márcia Aparecida Fernandes, Carlos Roberto Lopes, Fabiano Azevedo Dorça, and Luciano Vieira Lima. A stochastic approach for automatic and dynamic modeling of students learning styles in adaptive educational systems. *Informatics in Education-An International Journal*, (Vol11.2):191–212, 2012.
9. Cheryl Jones, Carla Reichard, and Kouider Mokhtari. Are students' learning styles discipline specific? *Community College Journal of Research & Practice*, 27(5):363–375, 2003.
10. Fabiano Azevedo Dorça, Luciano Vieira Lima, Márcia Aparecida Fernandes, and Carlos Roberto Lopes. A new approach to discover students learning styles in adaptive educational systems. *Revista Brasileira de Informática na Educação*, 21(01):76, 2013.
11. Fabiano A Dorça, Luciano V Lima, Márcia A Fernandes, and Carlos R Lopes. Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: An experimental analysis. *Expert Systems with Applications*, 40(6):2092–2101, 2013.
12. Curtis A Carver, RA Howard, and WD Lane. Addressing different learning styles through course hypermedia. *IEEE Transactions on Education*, 42(1):33–38, 1999.
13. Sabine Graf, Silvia Rita Viola, and T Leo Kinshuk. Representative characteristics of felder-silverman learning styles: An empirical model. In *Proceedings of the IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2006), Barcelona, Spain*, pages 235–242, 2006.