20

Backward Compatible Spatialized Teleconferencing based on Squeezed Recordings

Christian H. Ritz[1], Muawiyath Shujau[1], Xiguang Zheng[1], Bin Cheng[1], Eva Cheng[1,2] and Ian S Burnett[2]
[1]*School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong,*
[2]*School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia*

1. Introduction

Commercial teleconferencing systems currently available, although offering sophisticated video stimulus of the remote participants, commonly employ only mono or stereo audio playback for the user. However, in teleconferencing applications where there are multiple participants at multiple sites, spatializing the audio reproduced at each site (using headphones or loudspeakers) to assist listeners to distinguish between participating speakers can significantly improve the meeting experience (Baldis, 2001; Evans et al., 2000; Ward & Elko 1999; Kilgore et al., 2003; Wrigley et al., 2009; James & Hawksford, 2008). An example is Vocal Village (Kilgore et al., 2003), which uses online avatars to co-locate remote participants over the Internet in virtual space with audio spatialized over headphones (Kilgore, *et al.*, 2003). This system adds speaker location cues to monaural speech to create a user manipulable soundfield that matches the avatar's position in the virtual space. Giving participants the freedom to manipulate the acoustic location of other participants in the rendered sound scene that they experience has been shown to provide for improved multitasking performance (Wrigley et al., 2009).

A system for multiparty teleconferencing requires firstly a stage for recording speech from multiple participants at each site. These signals then need to be compressed to allow for efficient transmission of the spatial speech. One approach is to utilise close-talking microphones to record each participant (e.g. lapel microphones), and then encode each speech signal separately prior to transmission (James & Hawksford, 2008). Alternatively, for increased flexibility, a microphone array located at a central point on, say, a meeting table can be used to generate a multichannel recording of the meeting speech A microphone array approach is adopted in this work and allows for processing of the recordings to identify relative spatial locations of the sources as well as multichannel speech enhancement techniques to improve the quality of recordings in noisy environments. For efficient transmission of the recorded signals, the approach also requires a multichannel compression technique suitable to spatially recorded speech signals.

A recent approach for multichannel audio compression is MPEG Surround (Breebaart et al., 2005). While this approach provides for efficient compression, it's target application is loudspeaker signals such as 5.1 channel surround audio rather than microphone array recordings. More recently, Directional Audio Coding (DirAC) was proposed for both compression of loudspeaker signals as well as microphone array recordings (Pulkki, 2007) and in (Ahonen et al., 2007), an application of DirAC to spatial teleconferencing was proposed. In this chapter, an alternative approach based on the authors' Spatially Squeezed Surround Audio Coding (S^3AC) framework (Cheng et al., 2007) will be presented. In previous work, it has been shown that the S^3AC approach can be successfully applied to the compression of multichannel loudspeaker signals (Cheng et al., 2007) and has some specific advantages over existing approaches such as Binaural Cue Coding (BCC) (Faller et al., 2003), Parametric Stereo (Breebaart et al., 2005) and the MPEG Surround standard (Breebaart, et al., 2005). These include the accurate preservation of spatial location information whilst not requiring the transmission of additional side information representing the location of the spatial sound sources. In this chapter, it will be shown how the S^3AC approach can be applied to microphone array recordings for use within the proposed teleconferencing system. This extends the previous work investigating the application of S^3AC to B-format recordings as used in Ambisonics spatial audio (Cheng et al., 2008b) as well as the previously application of S^3AC to spatialized teleconferencing (Cheng et al., 2008a).

For recording, there are a variety of different microphone arrays that can be used such as simple uniform linear or circular arrays or more complex spherical arrays, where accurate recording of the entire soundfield is possible. In this chapter, the focus is on relatively simple microphone arrays with small numbers of microphone capsules: these are likely to provide the most practical solutions for spatial teleconferencing in the near future. In the authors' previously proposed spatial teleconferencing system (Cheng et al., 2008a), a simple four element circular array was investigated. Recently, the authors have investigated the Acoustic Vector Sensor (AVS) as an alternative for recording spatial sound (Shujau et al., 2009). An AVS has a number of advantages over existing microphone array types including their compact size (occupying a volume of approximately 1 cm^3) whilst still being able to accurately record sound sources and their location. In this chapter, the S^3AC will be used to process and encode the signals captured from an AVS.

Fig. 1 illustrates the conceptual framework of the multi-party teleconferencing system with N geographically distributed sites concurrently participating in the teleconference. At each site, a microphone array (in this work an AVS) is used to record all participants and the resulting signals are then processed to estimate the spatial location of each speech source (participant) relative to the array and to enhance the recorded signals that may be degraded by unwanted noise present in the meeting room (e.g. babble noise, environmental noise). The resulting signals are then analysed to derive a downmix signal using the S^3AC representing the spatial meeting speech. The downmix signal is an encoding of the individual speech signals as well as information representing their original location at the participants' site. The downmix could be a stereo signal or a mono signal. For a stereo (two channel) downmix, spatial location information for each source is encoded as a function of the amplitude ratios of the two channels; this requires no separate transmission of spatial location information. For a mono (single channel) downmix, separate information representing the spatial location of the sound sources is transmitted as side information. In either approach, the downmix signal is further compressed in a backwards compatible

approach using standard audio coders such as the Advanced Audio Coder (AAC) (Bosi & Goldberg, 2002). Since the application of this chapter is spatial teleconferencing, downmix compression is achieved using the extended Adaptive Multi-Rate Wide Band (AMR-WB+) coder (Makinen, 2005). This coder is chosen as it is one of the best performing standard coders at low bit rates for both speech and audio (Makinen, 2005) and is particularly suited to S^{3}AC. In Fig. 1, each site must unambiguously spatialise N-1 remote sites and utilizes a standard 5.1 playback system, however, the system is not restricted to this and alternative playback scenarios could be used (e.g. spatialization via headphones using Head Related Transfer Functions (HRTFs) (Cheng et al., 2001).

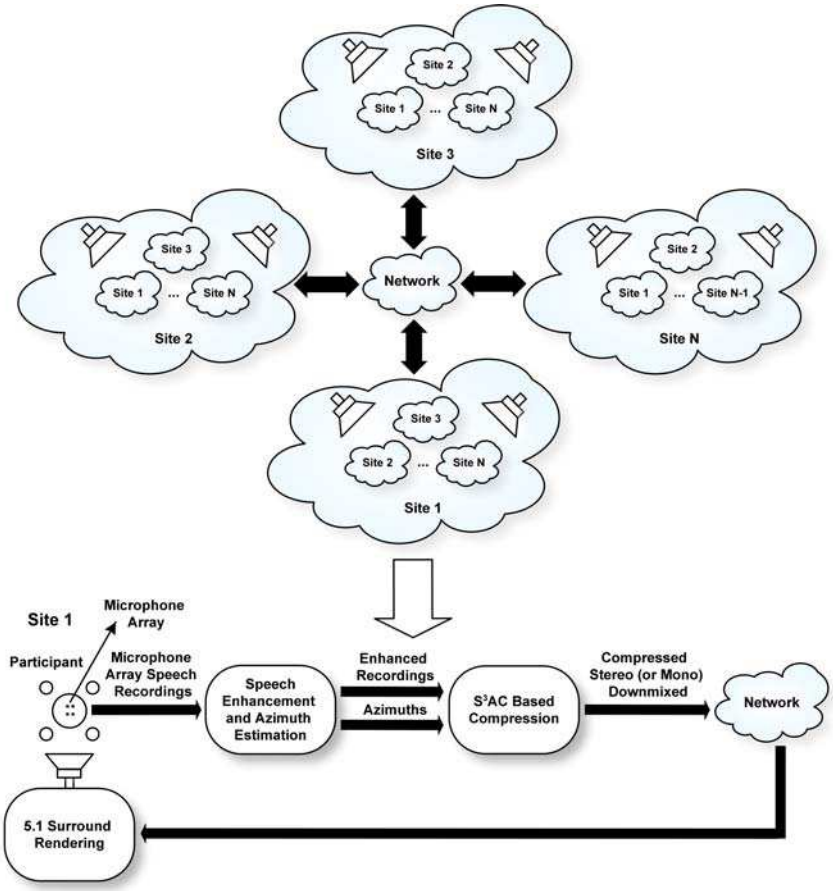


Fig. 1. Conceptual Framework of the Spatial Teleconferencing System. Illustrated are multiple sites each participating in a teleconference as well as a system overview of the S^{3}AC-based recording and coding system used at each site.

A fundamental principle of S^{3}AC is the estimation of the location of sound sources and this requires estimation of the location of sources corresponding to each speaker. In (Cheng et al., 2008a), the speaker azimuths were estimated using using the Steered Response Power

with PHAse Transform (SRP-PHAT) algorithm (DiBiase et al., 2001). This technique is suited to spaced microphone arrays such as the circular array presented in Fig. 1 and relies on Time-Delay Estimation (TDE) applied to microphone pairs in the array. In the current system, the AVS is a co-incident microphone array and hence methods based on TDE such as SRP-PHAT are not directly applicable. Hence in this work, source location information will be found by performing Directional of Arrival (DOA) estimation using the Multiple Signal Classification (MUSIC) method as proposed in (Shujau et al., 2009).
In this chapter two multichannel speech enhancement techniques are investigated and compared: a technique based on the Minimum Variance Distortionless Response (MVDR) beamformer (Benesty et al., 2008); and an enhancement technique based on sound source separation using Independent Component Analysis (ICA) (Hyvärinen et al., 2001). In contrast to existing work, these enhancement techniques are applied to the coincident AVS microphone array and results will extend those previously described in (Shujau et al., 2010).
The structure of this chapter is as follows: Section 2 will describe the application of S^{3}AC to the proposed teleconferencing system while Section 3 will describe the recording and source location estimation based on the AVS; Section 4 will describe the experimental methodology adopted and present objective and subjective results for sound source location estimation, speech enhancement and overall speech quality based on Perceptual Evaluation of Speech Quality (PESQ) (ITU-R P.862, 2001) measures; Conclusions will be presented in Section 4.

2. Spatial teleconferencing based on S3AC

In this section, an overview of the S^{3}AC based spatial teleconferencing system will first be presented followed by a detailed description of the transcoding and decoding stages of the system.

2.1 Overview of the system

Fig. 2 describes the high level architecture of the proposed spatial teleconferencing system based on S^{3}AC. Each site records one or more sound sources using a microphone array and these recordings are analysed to derive individual sources and information representing their spatial location using the source localisation approaches illustrated in Fig. 1 and described in more detail in Section 3. In this work, spatial location is determined only as the azimuth of the source in the horizontal plane relative to the array. In Fig. 2 sources and their corresponding azimuth are indicated as Speaker 1 + Azimuth to Speaker N + Azimuth.
The resulting signals from one or more sites are input to the S^{3}AC transcoder that processes the signals using the techniques to be described in Section 2.2 to produce a downmix signal that encodes the original soundfield information. The downmix signal can either be a stereo signal (labeled as S^{3}AC-SD in Fig. 2), where information about the source location is encoded as a function of the amplitude ratio of the two signals (see Section 2.2) or a mono-signal (labeled as S^{3}AC-SD in Fig. 2), where side-information is used to encode the source location information. In the implementation described in this work, the downmix is compressed using the AMR-WB+ coder, as illustrated in Fig. 2. This AMR-WB+ coder was chosen to provide backwards compatibility with a state-of-the-art standardised coder that has been shown to provide superior performance for speech and mixtures of speech and other audio at low bit rates (6 kbps up to 36 kbps), which is the target of this work.

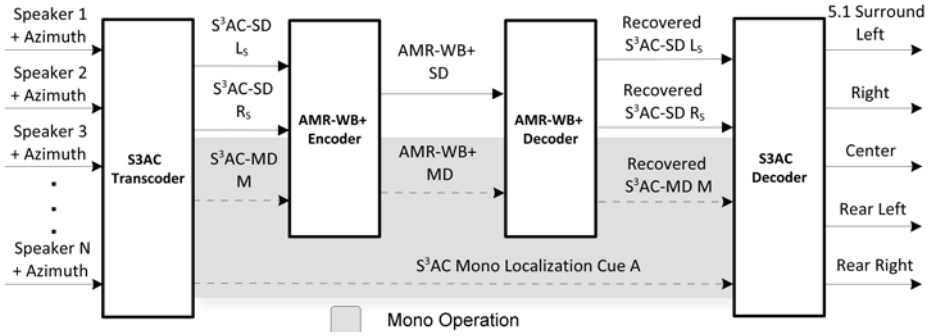


Fig. 2. High Level Architecture of the S³AC based teleconferencing system. S³AC-SD refers to the Stereo Downmix mode while S³AC-MD refers to the optional Mono Downmix mode. Speaker 1 to Speaker N refers to the recorded signals from one or more sites.

At the decoder, following decoding by the speech codec, the received downmix signals are analysed using the S³AC decoder described in Section 2.3 to determine the encoded source signals and information representing their spatial location. It should be noted that the spatial information represents the original location of each speaker relative to a central point at the recording site. The final stage is rendering of a spatial soundfield representing the teleconference, which is achieved using a standard 5.1 Surround Sound loudspeaker system (although alternative spatialization techniques may also be used due to the coding framework representing sound sources and their locations, which provides for alternative spatial rendering).

2.2 S³AC transcoder

An illustration of the S³AC transcoder is shown in Fig. 3 and consists of three main stages: Time-Frequency Transformation, Spatial Squeezing and Inverse Time-Frequency Transformation. Input to the S³AC transcoder are the speaker signals and corresponding azimuths of Fig. 2. Here, $s_{i,j}(n)$ and $\theta_{i,j}(n)$ are defined asthe speech source j and corresponding azimuth at site i, where $i=1$ to N and $j = 1$ to M_j and where N is the number of sites and M_j is the number of participants (unique speech sources) at each site.

In Fig. 3, this notation is used to indicate for site 1, signals representing the recorded sources and their corresponding azimuths. These signals are converted to the Fourier domain using a short time Fourier transform to produce the frequency domain signals $S_{i,j}(n,k)$, where n represents the time frame and k represents discrete frequency. Here, similar to the existing principle of S³AC, a separate azimuth is determined for each time-frequency component using the direction of arrival estimation approaches described in Section 3. While the azimuth is not expected to vary widely with frequency when a single participant is speaking, there will be variation when multiple participants are speaking concurrently; hence azimuths are denoted $\theta_{i,j}(n,k)$. This indicates that at each time and frequency there could be one or more speakers active at one or more sites.

The second stage of the S³AC transcoder is spatial squeezing, which assigns a new azimuth for the sound source in a squeezed soundfield. Conceptually, this involves a mapping of the source azimuth derived for the original 360° soundfield of the recording site to a new azimuth within a smaller region of a virtual 360° soundfield that represents all sources from all sites. This process can be described as:

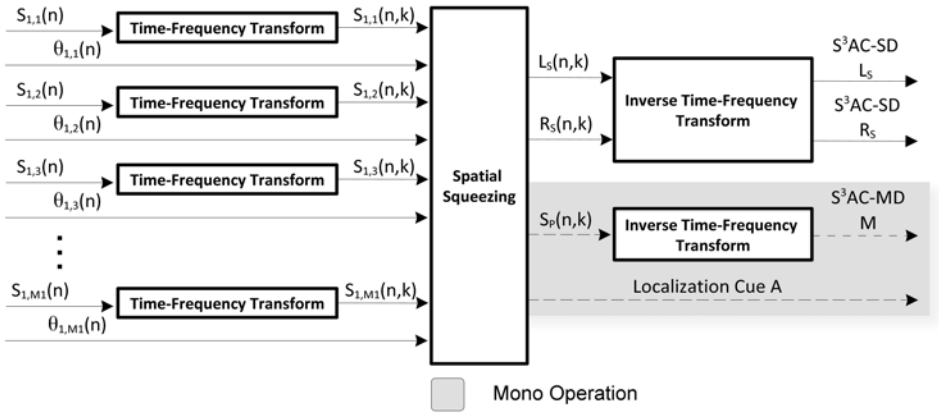


Fig. 3 S3AC Transcoder showing the encoding of multiple spatial speech signals and their azimuths as a time domain stereo (or optional mono) downmix signal.

$$\theta_{i,j}^{s}(n,k) = f\left(\theta_{i,j}(n,k)\right) \tag{1}$$

where f is a mapping function, which can be thought of as a quantization of the original azimuth to the squeezed azimuth. Examples of mapping functions for spatial audio compression are described in (Cheng et al., 2006). Here, a uniform quantization approach is adopted, whereby each azimuth is mapped to a squeezed azimuth equal to one of a possible $360/N$ quantized azimuths; conceptually, this divides the virtual soundfield into N equal regions, each representing one of the N remote sites. Following azimuth mapping, a downmix signal is created using one of two possible. Firstly, a stereo downmix can be created using the approach described by:

$$\begin{aligned} L_S(n,k) &= \frac{S_p(n,k)\cdot\left(\tan\varphi_d + \tan\theta_{p,s}(n,k)\right)}{\sqrt{2\tan^2\varphi_d + 2\tan^2\theta_{p,s}(n,k)}} \\ R_S(n,k) &= \frac{S_p(n,k)\cdot\left(\tan\varphi_d - \tan\theta_{p,s}(n,k)\right)}{\sqrt{2\tan^2\varphi_d + 2\tan^2\theta_{p,s}(n,k)}} \end{aligned} \tag{2}$$

where the left and right channel of the stereo signals, L_S and R_S, have an angular separation of $2\varphi_d$ and this approach encodes the azimuth as the ratio of the downmix signals and hence requires no separate representation (or transmission) of spatial information. In(2), $S_P(n,k)$ represents the primary spatial sound source corresponding to the active speech at a given time at frequency over all participants and sites. This is determined as the source with the highest magnitude using (3).

$$S_p(n,k) = \max_{i,j}\left(\left|S_{i,j}(n,k)\right|\right) \tag{3}$$

For non-concurrently speaking participants, this will correspond to the speech of the only person speaking. In the alternative mono-downmix approach (see Fig. 3), the downmix is

simply equal to the primary sound sources, $S_p(n,k)$. This approach requires separate representation (and transmission) of the azimuth information. For either downmix approach, the resulting signal is passed through an inverse time-frequency transform to create a time-domain downmix for each frame. This is the final stage of Fig. 3. The output of the transcoder is then fed to the AMR-WB+ encoder block of Fig. 3 prior to transmission..

2.3 S^{3}AC decoder

The S^{3}AC decoder block of Fig. 2 is illustrated in more detail in Fig. 4. Following speech decoding, the resulting received downmix signals are converted to the frequency domain using the same transform as applied in the S^{3}AC transcoder. These signals are then fed to the spatial repanning stage of Fig. 4. In the stereo-downmix mode, spatial repanning applies inverse tangent panning to the decoded stereo signals $\hat{R}_s(n,k)$ and $\hat{L}_s(n,k)$ to derive the squeezed azimuth of the time-frequency virtual source, $\hat{\theta}_{p,s}(n,k)$, using (4):

$$\hat{\theta}_{p,s}(n,k) = \arctan\left(\frac{\hat{L}_s(n,k) - \hat{R}_s(n,k)}{\hat{L}_s(n,k) + \hat{R}_s(n,k)} \cdot \tan\varphi_d \right) \tag{4}$$

The original azimuth $\hat{\theta}^s_{i,j}(n,k)$ of this virtual source is then recovered using:

$$\hat{\theta}^s_{i,j}(n,k) = f^{-1}\left(\hat{\theta}_{p,s}(n,k)\right) \tag{5}$$

In Equation (4), f^{-1} represents the inverse azimuth mapping function used in Equation (1). Following decoding of the original azimuth of the primary source, an estimate of the primary source $\hat{S}_p(n,k)$ is obtained using Equation (2) and the estimated primary source azimuths and decoded downmix signals.

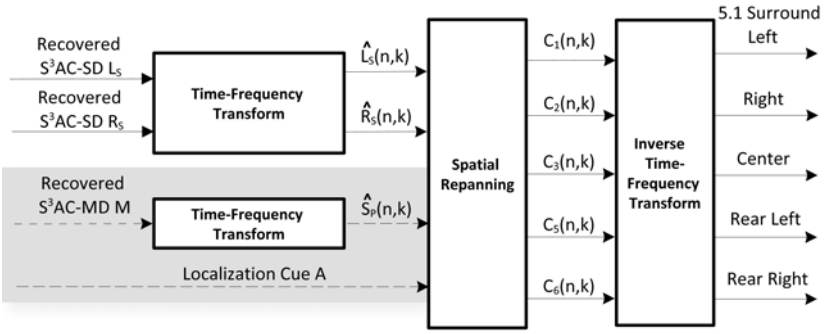


Fig. 4. S^{3}AC Decoder illustrating the processing of time domain signals recovered by the AMR-WB+ decoder to produce time-domain loudspeaker signals for reproduction of the spatial teleconference audio at each site.

The final rendering stage of the spatial re-panning is dependent on the desired playback system at each site. Illustrated in Fig. 4 is the scenario whereby reproduction at each site is achieved using a standard 5.1 channel Surround Sound loudspeaker system and utilizing all

channels other than the low frequency effect channel. In this scenario, the estimated primary sources are amplitude panned to the desired location using two channels of the 5 channel system. This can be achieved by re-applying equation (2) using the azimuthal separation of the chosen two channels in the playback system and the estimated primary source azimuth. The output of this stage is a set of frequency-domain loudspeaker channel signals and the final step is to apply an inverse time-frequency transform to obtain the time-domain loudspeaker signals. Other reproduction techniques are also possible (e.g. binaural reproduction using HRTF processing (Cheng, 2008b). Due to the preservation of the original spatial location of each participant at each site, rendering could include accurate spatialization for virtual recreation of remote participants (e.g. for correct positioning of speech signals to correspond with the videoed participants). Alternatively, positioning could be achieved interactively at each site such as described in (Kilgore et al., 2003). In this chapter the primary focus is to ensure the perceptual quality resulting from decoding of each of the received spatial speech signals and hence further discussion on spatial rendering is not included.

3. An AVS for spatial teleconferencing

3.1 Overview of the AVS

An AVS consists of three orthogonally mounted acoustic particle velocity sensors and one omni-directional acoustic pressure microphone, allowing the measurement of scalar acoustic pressure and all three components of acoustic particle velocity (Hawkes & Nehorai, 1996; Lockwood & Jones, 2006). A picture of the AVS used in this work is shown in Fig. 5. Compared to linear microphone arrays, AVS's are significantly more compact (typically occupying a volume of 1 cm^3) (Hawkes & Nehorai, 1996; Lockwood & Jones, 2006; Shujau et al., 2009) and can be used to record audio signals in both the azimuth and elevation plane. Fig. 2 presents a picture of the AVS developed in (Shujau et al., 2009). The acoustic pressure and the 2D (x and y) velocity components of the AVS can be expressed in vector form as:

$$\mathbf{s}(n) = [o(n), x(n), y(n)]^T \tag{6}$$

In (6), **s**(n), is the vector of recorded samples, where $o(n)$ represents the acoustic pressure component measured by the omni-directional microphone and $x(n)$ and $y(n)$ represent the outputs from two gradient sensors that estimate the acoustic particle velocity in the x and y direction, relative to the microphone position. For the gradient microphones, the relationship between the acoustic pressure and the particle velocity is given by Equation (7) (Shujau et al., 2009):

$$[x(n), y(n)] = g(p(n) - p(n - \Delta n))u \tag{7}$$

Equation (7) assumes a single primary source, where g represents a function of the acoustic pressure difference and:

$$\mathbf{u} = \begin{bmatrix} \cos\theta_{i,j} & \sin\theta_{i,j} \end{bmatrix}^T \tag{8}$$

is the source bearing vector with $\theta_{i,j}$ representing the azimuth of the single source relative to the microphone array (Shujau et al., 2009).

3.2 Direction of arrival estimation of speech sources using the AVS

Directional information from an AVS can be extracted by examining the relationship between the 3 microphone channels. Accurate Direction of Arrival (DOA) estimates are dependent upon placement of the microphones, the structure that holds the microphones and the polar patterns generated by each microphone. A design that results in highly accurate DOA estimation using the Multiple Signal Classification (MUSIC) method of Schmidt (Schmidt, 1979) was presented in (Shujau et al., 2009) and is adopted here.

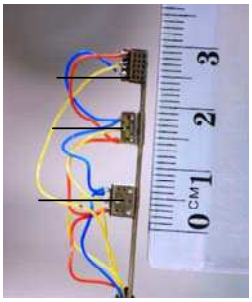


Fig. 5. The Acoustic Vector Sensor (AVS) used for recording of the spatial teleconference at each site.

The MUSIC algorithm allows for the estimation of the source DOA using the eigenvalues and eigenvectors of the covariance matrix formed from the recorded signals (Manolakis et al. 2005; Schmidt, 1979). The covariance matrix formed from the recorded signals is described in Equation (9), where L represents the number of samples used to find the covariance matrix (in this work, L corresponds to a single frame of 20 ms duration).

$$R(n) = \frac{1}{L}\sum_{n=1}^{L} \mathrm{Re}\left\{ s(n)s^{*}(n) \right\} \tag{9}$$

The MUSIC algorithm is then used to estimate the azimuth of source j at site i, $\theta_{i,j}$, using Equation (10).

$$\theta_{i,j} = \min_{\theta}\left[P(\theta) = \frac{1}{\sum \left| \mathbf{V}^{H}\mathbf{h}(\theta) \right|^{2}} \right] \tag{10}$$

where $\mathbf{V}$ is the smallest eigenvector of the covariance matrix R from (9) and $\mathbf{h}(\theta)$ is the steering vector for the AVS and $\theta \in$ (-π, π). Assuming sources are only in the 2D plane, relative to the microphone array, , the steering vector can be described as a function of the azimuth as (Manolakis et al. 2005; Schmidt, 1979):

$$\mathbf{h}(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) & 1 \end{bmatrix} \tag{11}$$

which is formed from the x and y components of Equation (6) and where 1 represents the omni-directional microphone.

3.3 Enhancement of AVS recordings

Speech enhancement for the AVS is achieved using two methods. The first method uses Independent Component Analysis (ICA) (Hyvärinen et al., 2001) while the second method uses the Minimum Variance Distortionless Response (MVDR) beamformer (Benesty et al., 2008).

3.3.1 Enhancement via ICA

The traditional ICA model applied to a multichannel speech recording assumes that microphone frequency responses for each channel are the same and that the mixing matrix is a result only of the acoustic transfer function. However, for the AVS, the microphones have directional polar responses and an approach for ICA for the AVS was previously described in (Shujau et al., 2010). This work applies ICA to recordings of the acoustic pressure gradients. In ICA, the aim is to separate a set of mixed signals into signal representing one or more independent sources. Here, the case for two source signals and 3 microphones is first considered. The recorded signals can be modeled using the mixing model:

$$\hat{\mathbf{s}}(n) = \sum_{j=0}^{2} \mathbf{A_k} \mathbf{s}_j(n) \tag{12}$$

In Equation (12), $\hat{\mathbf{s}}(n)$ represents a model of the recorded signals $\mathbf{s}(n)$ of equation (6), and $\mathbf{s}_j(n) = [s_1(n), s_2(n)]^T$ represents the vector of source signal samples and $\mathbf{A_k}$ represents the convolutive mixing matrices, each of size 3×2. In the case where there is only one speaker in the presence of diffuse noise, the output components following ICA will be the primary speech source as well as residual noise signals. Here, for anechoic recordings, ICA was implemented using the well known FastICA implementation (Hyvärinen et al., 2001) while reverberant recordings were processing using a convolutive FastICA algorithm (Douglas et al., 2005).

3.3.2 Ehancement via MVDR

The MVDR Beamformer is the most widely used beamformer for microphone arrays. The expected outcome of any beamformer for speech is to combine the sensor signals in such a way that the desired speech signal is preserved or enhanced while the interfering signals are reduced without introducing any distortion. In this work a frequency domain MVDR beamformer is implemented. The MVDR beamformer is formed by choosing the coefficients of the filter $\mathbf{w}$ such that output power $E\left[Z^2\right] = \mathbf{w}^T \mathbf{R}(n,k)\mathbf{w}$ is minimized without introducing any distortion to the source signal (Benesty et al., 2008) where $\mathbf{R}$ is the covariance matrix of Equation (9) in the frequency domain. For each 20 ms frame, an FFT of 1024 samples is found using a Hamming window with an overlap of 50 %. The frequency domain samples are represent by the components of a vector $\mathbf{S}(n,k) = [x(n,k)\ y(n,k)\ o(n,k)]$ where n is the

sample number and k is the frequency bin. The F = 32 most recent frames are buffered and the covariance matrix $\mathbf{R}(n,k)$ of the vector $\mathbf{S}(n,k)$ is found as (Lockwood et al., 2004):

$$\mathbf{R}(n,k) = \begin{bmatrix} \frac{c}{F}\sum_{l=0}^{F-1} x(m_l,k)^* x(m_l,k) & \frac{1}{F}\sum_{l=0}^{F} x(m_l,k)^* y(m_l,k) & \frac{1}{F}\sum_{l=0}^{F} x(m_l,k)^* o(m_l,k) \\ \frac{1}{F}\sum_{l=0}^{F} y(m_l,k)^* y(m_l,k) & \frac{c}{F}\sum_{l=0}^{F} y(m_l,k)^* y(m_l,k) & \frac{1}{F}\sum_{l=0}^{F} y(m_l,k)^* o(m_l,k) \\ \frac{1}{F}\sum_{l=0}^{F} o(m_l,k)^* x(m_l,k) & \frac{1}{F}\sum_{l=0}^{F} o(m_l,k)^* y(m_l,k) & \frac{c}{F}\sum_{l=0}^{F} o(m_l,k)^* o(m_l,k) \end{bmatrix} \quad (13)$$

Where m_l=n-lL and c = 1.03 which is regularization constant to help avoid matrix singularity and * represents complex conjugate. The covariance matrix is updated every 16 frames. The MVDR filter is expressed as (Lockwood et al., 2004):

$$\mathbf{w_k} = \frac{\mathbf{R}_{ky}^{-1}\mathbf{h}}{\mathbf{h}^T\mathbf{R}_{ky}^{-1}\mathbf{h}} \quad (14)$$

where $\mathbf{h}$ is the steering vector of Equation (10) and and the optimization constraints for each frequency band are described as:

$$\min_{\mathbf{w}} \mathbf{w}^T\mathbf{R}_{ky}\mathbf{w} \qquad \text{subject to} \qquad \mathbf{w}^T\mathbf{h} = 1 \quad (15)$$

The output of the beamformer for each frequency band k is given by (Lockwood et al., 2004):

$$Z_k = \mathbf{w_k^H Y_k} \quad (16)$$

The time domain output is obtained by determining the inverse FFT and performing overlap add reconstruction.

4. Experimental evaluation

In this section, results are presented evaluating the source localisation performance, enhancement performance and overall speech quality using the AVS and S³AC based spatial teleconferencing system.

4.1 Experimental evaluations

An experimental rig was created, where the AVS was mounted on a custom built rotating platform to allow positioning of the microphones relative to the source. Sound sources were produced by self powered speakers (Genelec 8020A) located at 1 m from the array. For source localization experiments, a series of monotone signals each 2 seconds long and of equal energy were recorded in an anechoic room with frequencies ranging from 100 Hz to 10 kHz. The recordings were made with the microphone rotated in 5˚ intervals corresponding to sources located at azimuths ranging from 0˚ to 90˚, hence covering a full quadrant in the x-y plane. Recordings were also made using speech sources in both anechoic and reverberant conditions (with RT_{60} of 30ms), using 12 speech sentences (six male and six

female) from the IEEE speech corpus (IEEE Subcommittee, 1969). Each sentence is 10 s long with 1s of silence at the start and end. Five 10 s segment noise sources are utilised: babble; recordings of factory floor; background noise from a moving vehicle; white; and pink noise, which were taken from an existing database (Institute for Perception-TNO, 1990). Diffuse noise was simulated using 4 loudspeakers located at equal distances on a circle surrounding the array. Recordings were made of a single target speech source in the presence of diffuse noise as well as one or two speech interferers as the noise source. The recordings were sampled at a rate of 48 kHz and two different Signal-to-Noise Ratio (SNR) levels of 0 dB and 20 dB. For source localization experiments, recordings were also made with a Uniform Linear Array (ULA) with a similar number of capsules to the AVS and a SoundField ST-250 microphone (SoundField) and the MUSIC algorithm was also used for the DOA estimation for these recordings. The SoundField microphone was chosen as it provides a direct comparison with an existing co-located microphone array.

4.2 Sound source location estimation

To investigate the performance of the AVS for estimating the source location, a series of experiments were conducted. In these experiments, recordings using the AVS were processed using the MUSIC algorithm to estimate source directions. The source localization error was measured using the Average Angular Error (AAE), defined as the average error over all frames tested between the true and estimated DOA. For monotone sources in anechoic environments, as shown in Fig. 6, DOA estimates obtained from the AVS were found to have an average error of less than 2° for a range of source frequencies, compared with average errors of more than 4.5° for the ULA. In addition to the monotone sources, experiments were carried with speech sources recorded in the presence of diffuse noise. For these experiments, the reverberant recordings were considered rather than the easier scenario of DOA estimation in anechoic conditions of Fig. 6. Further, results were compared with the SoundField microphone rather than the inferior ULA. The results from these experiments (Fig. 7) show that the average error produced by the AVS for localising a speech source in diffuse noise for reverberant conditions is approximately 1.6^0 compared to that of the SoundField Microphone which is 5^0.

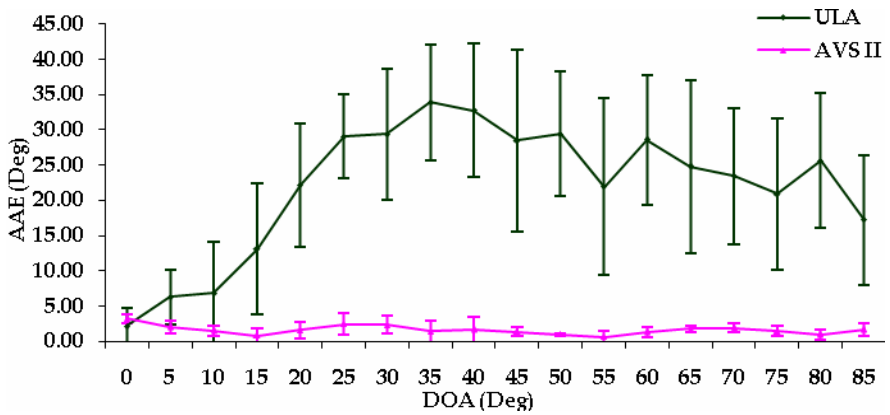


Fig. 6. Average Angular Error (AAE) for theDOA estimated for a series of tone sources with frequencies ranging from 1-10 kHz using both the AVS and the ULA. Recordings were made in anechoic conditions.

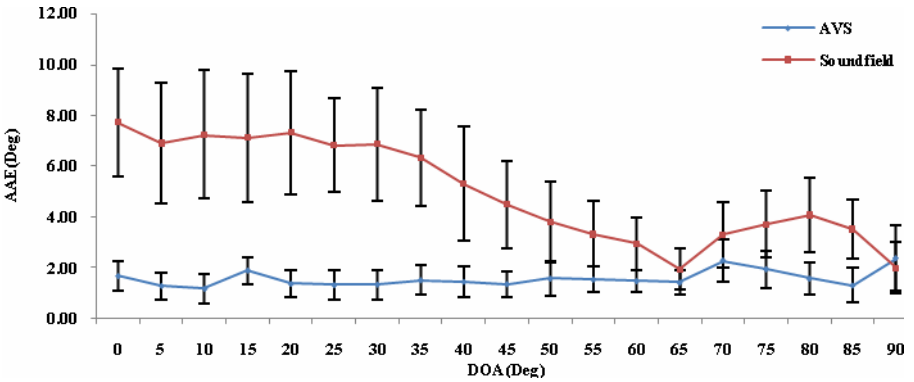


Fig. 7. Average Angular Error (AAE) for theDOA estimated for for recordings of speech from teh AVS and the SoundField microphones. Reocrdings were made in reverberant conditions.

4.3 Evaluation of speech enhancement

The results presented in this section are for two multi channel speech enhancement algorithms for the AVS, namely the ICA and MVDR beamformer. Both enhancement algorithms were used to process the recorded speech databases described in Section 4.1. The outputs from the enhancement algorithms are low pass filtered and down sampled to 16 kHz and then evaluated using the ITU-PESQ software (ITU-R P.862, 2001). When using PESQ, each output from ICA was compared with the original clean source signal to give a Mean Opinion Score for Listening Quality (MOSLQO) (Ma et al., 2009); the highest MOSLQO corresponds to the target source. A difference MOSLQO is generated by subtracting the MOSLQO of an omni-directional recording of the mixed sources (used as the reference) from the highest MOSLQO of the ICA outputs (Ma et al., 2009).

Results in Fig. 8 are for a speech source with both speech and diffuse noise as the interferer in anechoic conditions. The results show that on average when the recordings are enhanced with ICA there is an average improvement in MOSLQO of approximately 0.9 for diffuse noise as interferer and approximately 1.7 for speech as the interferer. In contrast, results for

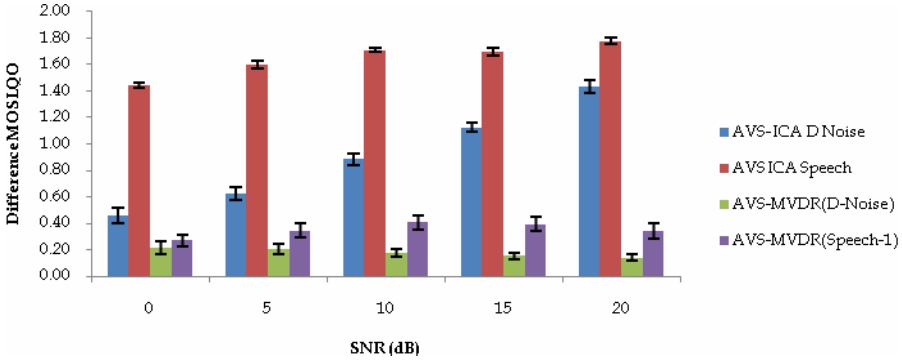


Fig. 8. Difference MOS Vs SNR in Anechoic Conditions

the MVDR based enhancement approach show a MOSLQO improvement of 0.2 and 0.4, respectively, for the diffuse noise and speech interferers. Hence, ICA shows a MOSLQO improvement of 0.7 and 1.3 over MVDR for diffuse noise and speech interferers, respectively.

Results for the reverberant case are shown in Fig. 9, where the difference MOSLQO for ICA is 0.7 for speech as the interferer and 0.5 for diffuse noise as interferer. In contrast, the MVDR enhancer results in an improved MOSLQO for both speech and noise interferers of 0.1. These results show that the ICA based enhancer is superior to the MVDR based enhancer in both anechoic and reverberant environments and for both diffuse and speech noise sources.

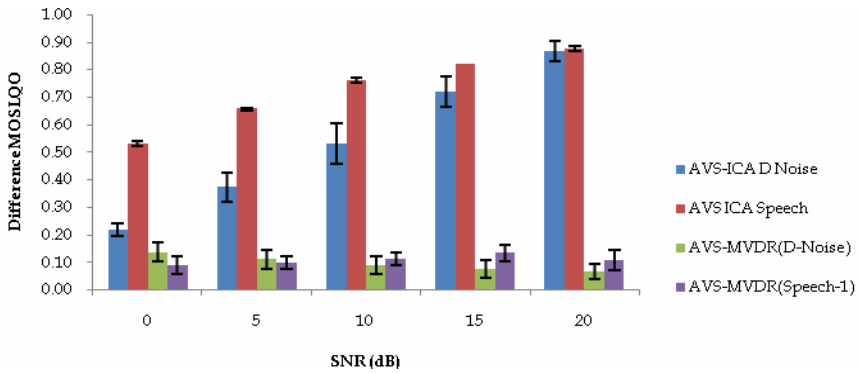


Fig. 9. Difference MOS Vs SNR in reverberant conditions.

4.4 Estimation of overall speech quality

To investigate the performance within the proposed spatial teleconferencing system, the recorded database of Section 4.1 was encoded through the proposed teleconferencing system including AMR-WB+ encoding of the downmix signals. The PESQ measure was used to analyse the resulting quality of the decoded signals that are the output of the proposed system. For the PESQ measures, the original clean sources were used as the reference. The AMR-WB+ coder was operated at each of the possible 31 bit rates ranging from 6 kbps to 36 kbps in increments of 1 kbps.

The first set of results for clean speech, where speech sources did not overlap in time, are shown in Fig. 10. The purpose of this test is to verify that the S^{3}AC coding framework does not introduce significant distortion additional to that introduced by the downmix compression. The results of Fig. 10 confirm that this is the case, with a gradual increase in PESQ as the bit rate of the AMR-WB+ coder increases. These results agree with existing results for the AMR-WB+ codec (Makin et al. 2005).

The second set of results shown in Fig. 11 is for recordings in the presence of diffuse noise with an SNR of 0 dB. The results in Fig. 11 (a) are for anechoic recordings where PESQ results have been averaged across those obtained for each of the five noise types and error bars represent 95 % statistical confidence intervals. Curves with blue asterisks represent results for PESQ of the decoded outputs from AMR-WB+ compared with the original clean

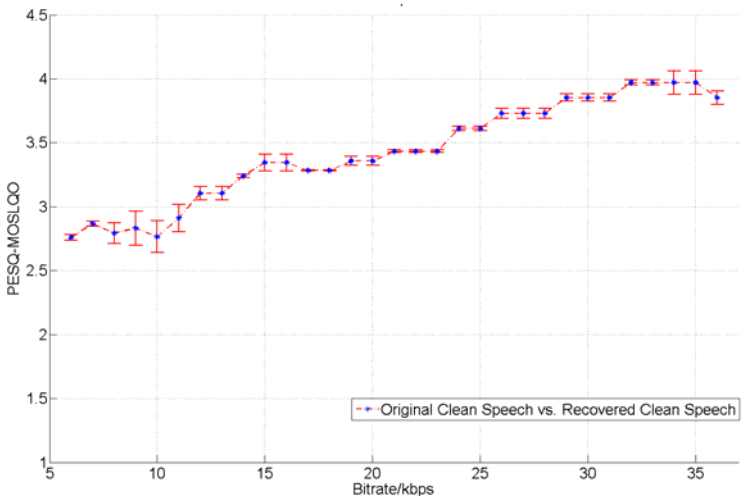
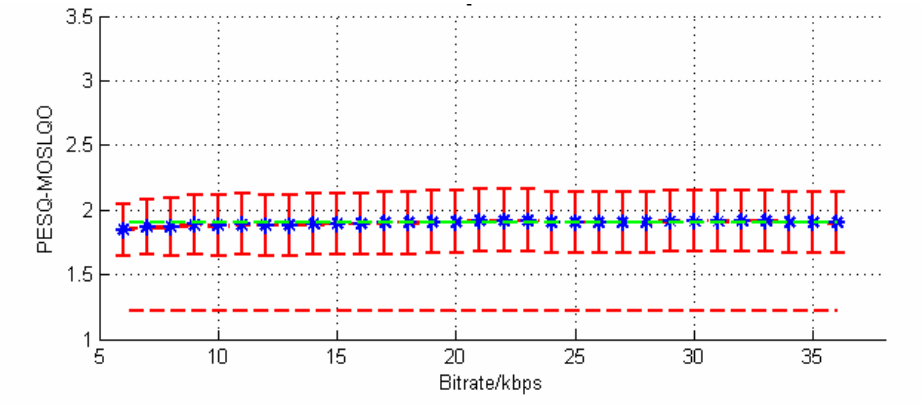


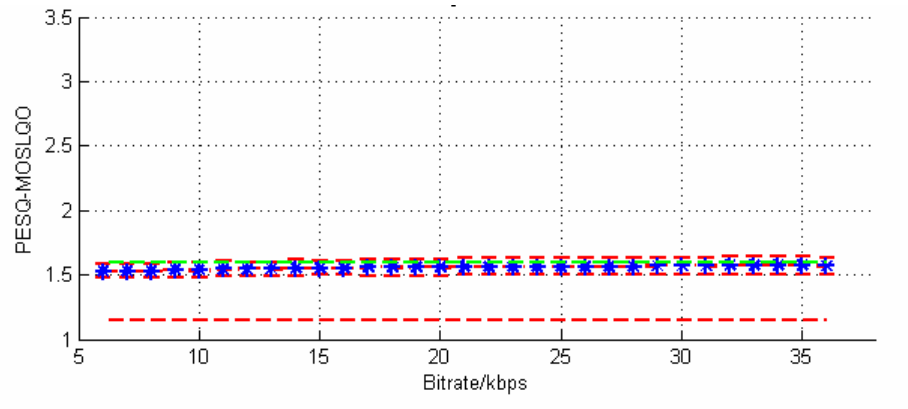
Fig. 10. PESQ results for original clean speech coded using the proposed spatial teleconferencing system including downmix compression via AMR-WB+.

speech. These results show a slight increase in PESQ over the bit rates tested with the average PESQ over all bit rates approximately 1.8. The green dashed line represents the upper limit for when no AMR-WB+ coding is applied to the enhanced recordings and it can be seen that the PESQ is statistically equivalent to the PESQ obtained after AMR-WB+ coding at the highest bit rates. It is proposed that this result is due to the high noise level present and hence further degradation caused by the speech coder does not dramatically reduce the resulting PESQ. The red dashed curve represents the results obtained when coding the non-enhanced recordings (taken as the omni-directional microphone output of the AVS) with the AMR-WB+ coder. As can be seen, the average PESQ is approximately 1.2, which is 0.6 less than results obtained when applying enhancement prior to encoding. The results for echoic recordings display similar trends to those for anechoic recordings, with PESQ results on average 0.4 higher for enhanced recordings compared with those for non-enhanced signals. On average, the PESQ results for enhanced recordings are 0.3 lower for echoic recordings compared to anechoic recordings.

Figure 12 shows results for anechoic and echoic recordings in the presence of noise where the SNR is 20 dB. In anechoic conditions (Fig. 12 (a)), the PESQ results are on average 2.6 for the enhanced signals decoded by AMR-WB+, which is an approximate 1.4 higher MOS prediction than for non-enhanced recordings. In echoic conditions (Fig. 12 (b)), average PESQ scores are approximately 2.2, which are is an approximate 0.9 increase in estimated MOS compared to non-enhanced signals. For bit rates of 14 kbps and above, results are statistically similar to those obtained when no speech coding is applied to the enhanced recordings. Compared with results for an SNR of 0 dB, the PESQ results for an SNR of 20 dB are approximately 0.8 higher. This result is to be expected due to the reduced level of noise present in the 20 dB SNR condition.



(a)



(b)

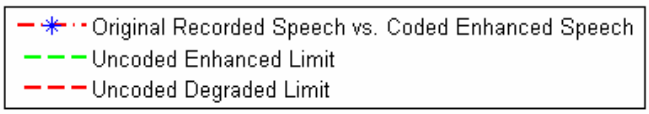
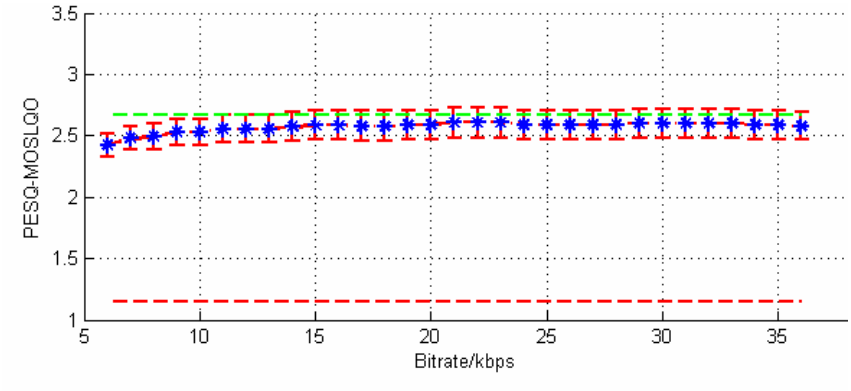
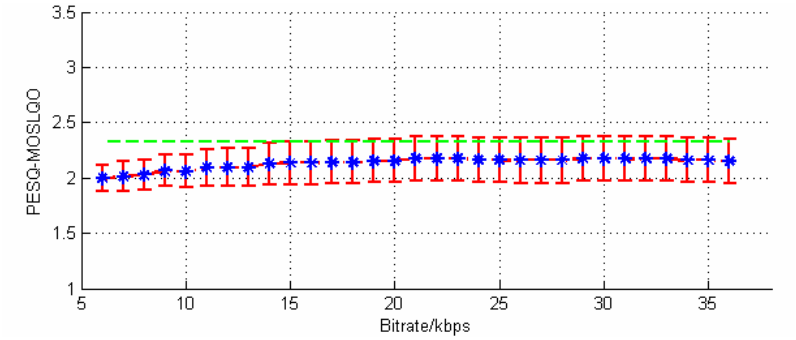


Fig. 11. Average results for recordings in diffuse noise across all noise sources for an SNR of 0 dB (a) Anechoic recordings. (b) Echoic recordings. Blue asterisk curves: results for PESQ of the decoded outputs from AMR-WB+ compared with the original clean speech. Green dashed curves: results when no AMR-WB+ coding is applied to the enhanced recordings. Red dashed curves: results obtained when coding the non-enhanced recordings.



(a)



(b)

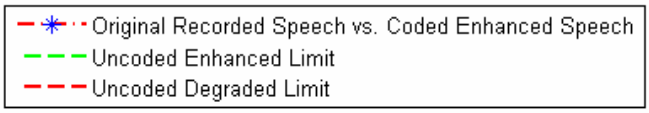


Fig. 12. Average results for recordings in diffuse noise across all noise sources for an SNR of 20 dB. (a) Anechoic recordings. (b) Echoic recordings. Blue asterisk curves: results for PESQ of the decoded outputs from AMR-WB+ compared with the original clean speech. Green dashed curves: results when no AMR-WB+ coding is applied to the enhanced recordings. Red dashed curves: results obtained when coding the non-enhanced recordings.

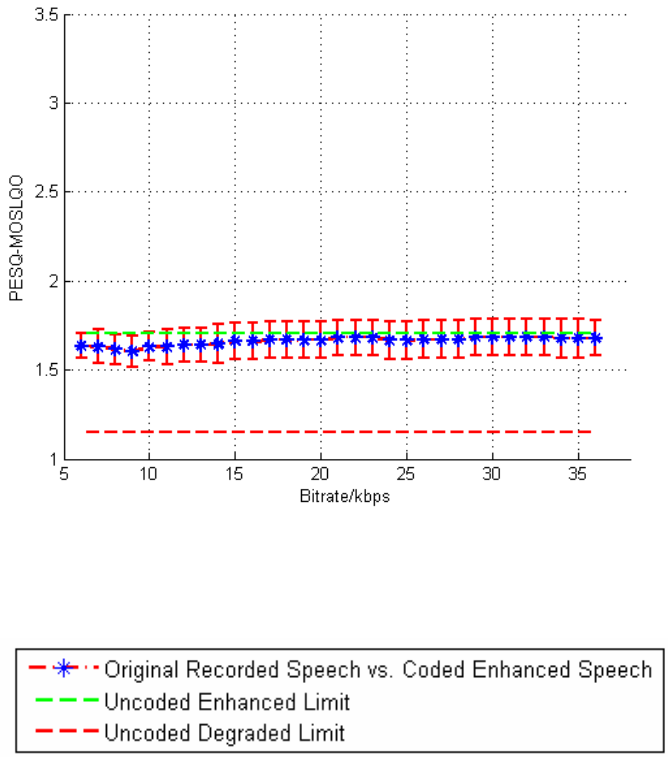


Fig. 13. PESQ results for recordings of two simultaneous speakers. Blue asterisk curves: results for PESQ of the decoded outputs from AMR-WB+ compared with the original clean speech. Green dashed curves: results when no AMR-WB+ coding is applied to the enhanced recordings. Red dashed curves: results obtained when coding the non-enhanced recordings.

The final set of results in Fig. 13 is for recordings of two simultaneous speech sources of equal power (SNR of 0 dB) separated by an angle of 45˚ and at a distance of 1 m from the microphone array. Here, one speech signal is treated as the desired source and the other as the interfering noise source. Results are for the PESQ of the output of the ICA enhancement that has been compressed and decoded using AMR-WB+, similar to the diffuse noise experiments. This shows that the enhancement results in an approximate 0.5 increase in estimated MOS for all bit rates tested. The PESQ results are also statistically similar to those obtained without AMR-WB+ coding of the ICA output.

5. Conclusion

This chapter has presented an approach to efficient compression for spatialized teleconferencing based on the concept of spatial squeezing of microphone array recordings of speech. Recordings were made using a collocated microphone array known as an AVS. Through encoding estimates of individual speech sources and information representing their spatial location, the proposed framework provides a flexible approach to the spatial rendering of the teleconference at each participants site. Results were presented confirming the accurate prediction of spatial sound sources though processing of the AVS recordings using the MUSIC algorithm. The approach results in a stereo or mono downmix signal representing the entire teleconference, which can then be efficiently compressed and transmitted using existing standard speech coders such as AMR-WB+. Hence, this provides for backward compatibility with existing speech coding and transmission systems.

Results were also presented demonstrating the performance of multichannel speech enhancement using a sound source separation inspired approach based on ICA. Predictions of subjective quality using PESQ showed that ICA-based enhancement results in a significant improvement in the predicted MOS compared to those obtained using the existing MVDR-based speech enhancer designed for microphone arrays. Results were also presented illustrating the performance in terms of PESQ when encoding the signals obtained from ICA enhancement using the proposed spatial teleconferencing system. These results show that the proposed approach does not introduce significant degradation in PESQ when compared with the PESQ obtained without encoding through the spatial teleconferencing compression system. Future work will focus on determining solutions to further enhancing the recorded signals when two or more participants are speaking simultaneously as well as improved methods for speech enhancement in low SNR conditions.

6. References

Ahonen, J.; Pulkki, V.; Lokki, T. (2007). "Teleconference Application and BFormat Microphone Array for Directional Audio Coding," Proc. AES 30th Int. Conf: Intelligent Audio Environments, March 2007.

Baldis, J. J. (2001). "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," Proc. ACM SIGCHI Conference on Human factors in Computing Systems, pp.166-173, Washington, USA, March 2001.

Benesty, J.; Chen J.; Huang, Y. (2008). "Microphone Array Signal; Processing," Springer, Berlin.

Bosi, M.; Goldberg, R.E. (2002). Introduction to Digital Audio Coding and Standards, Springer, ISBN:1-4020-7357-7.

Breebaart, J., et al. (2005a). "Parametric Coding of Stereo Audio", EURASIP Jour. Applied Signal Proc., 1305-1322, Sep. 2005.

Breebaart, J. et al. (2005b). "MPEG Spatial Audio Coding/MPEG Surround: Overview and Current Status", in Proc. 119th AES Convention, New York, USA, Oct., 2005.

Cheng, C.I.,; Wakefield, G.H. (2001). Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency and Space, J. Audio Eng. Soc., 49(4):231-249, Apr. 2001.

Cheng, B.; Ritz, C.; Burnett, I. (2006). "Squeezing the Auditory Space: A New Approach to Multi Channel Audio Coding", Advances in Multimedia Information Processing – PCM2006, *Proceedings of the 7th Pacific-Rim Conference on Multimedia (PCM2006)*, Hangzhou, China, Nov. 2-4, 2006, Lecture Notes in Computer Science 4261, pp. 572 - 581, Springer-Verlag, 2006.

Cheng, B.; Ritz, C.; Burnett, I. (2007). "Principles and Analysis of the Squeezing Approach to Low Bit Rate Spatial Audio", *Proc. 2007 IEEE International Conf. on Acoustic, Speech and Signal Processing (ICASSP2007)*, Volume 1, Pages 13-16, Apr. 2007.

Cheng, Eva; Cheng, Bin; Ritz, Christian; Burnett, Ian S. (2008a). "Spatialized Teleconferencing: Recording and 'Squeezed' Rendering of Multiple Distributed Sites", *Proc. Australasian Telecommunication Networks and Applications Conference*, Pages 441 – 416, Dec. 2008.

Cheng, B. Ritz C. H.; Burnett, I. S. (2008b). "Binaural reproduction of spatially squeezed surround audio," *Proc. ICSP 2008: 9th International Conference on Signal Processing*, (Beijing, China), 2008, pp. 506-509.

Cheng, B.; Ritz, C.; Burnett, I. (2008c). "A Spatial Squeezing Approach to Ambisonics Audio Compression", Proc. 2008 IEEE International Conf. on Acoustic, Speech and Signal Processing (ICASSP 2008), Las Vegas, USA, Mar. 2008.

Cheng, B.; Ritz, C.; Burnett, I. (2009). "Spatial Audio Coding by Squeezing: Analysis and Application to Compressing Multiple Soundfields", Proc. EUSIPCO 2009, Glasgow, Scotland, p. 909-913, 24-28 August 2009.

DiBiase, J. H.; Silverman, H. F.; Brandstein, M. S. (2001). "Robust localization in reverberant rooms,", Microphone Arrays: Techniques and Applications, M. Brandstein and D. Ward, Eds., Berlin: Springer- Verlag, 2001, pp. 157–180.

Douglas, S. C.; Sawada, H.; Makino, S. (2005). "A spatio-temporal fastICA algorithm for separating convolutive mixtures," IEEE ICASSP05, Vol.5, pp 165-168, March 2005.

Evans, M. J.; Tew A. I.; J.; Angus, A. S. (2000). "Perceived performance of loudspeaker-spatialized speech for teleconferencing," Journal of the Audio Engineering Society, vol. 48, no9, pp. 771-785.

Faller, C.; Baumgarte, F. (2005). "Binaural Cue Coding – Part II: Schemes and Applications", IEEE Trans. on Speech and Audio Proc., vol.11, No.6, Nov., 2003.

Hawkes, M.; Nehorai, A. (1996). "Bearing Estimates With Acoustic Vector Sensor Arrays", American Institute of Physics Con. Proc. Vol: 386, pp 345-358, April 1996.

Hyvärinen, A.; Karhunen, J.; Oja, E. (2001). "Independent Component Analysis," John Wiley & Sons Inc, New York.

IEEE Subcommittee (1969). IEEE Recommended Practice for Speech Quality Measurements", IEEE Trans. Audio and Electro-acoustics, AU-17(3), 225-246.

ITU-R P.862 (2001). ITU-R Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs".

Institute for Perception-TNO (1990). Noise Data, The Signal Processing Information Base (SPIB), Soesterberg, The Netherlands. Available online: http://spib.rice.edu/spib/select_noise.html

James, B. St.; Hawksford, M. O. J. (2008). "Corpuscular Streaming and Parametric Modification Paradigm for Spatial Audio Teleconferencing", Journal of the Audio Engineering Society, Volume 56 Issue 10 pp. 823-843, October 2008.

Kilgore, R.; Chignell, M.; Smith, P. (2003). "Spatialized audioconferencing: what are the benefits?", Proc. 2003 IBM Conference of the Centre for Advanced Studies on Collaborative Research, pp. 135-144, Ontario, Canada.

Lockwood, M. E.; Jones, D. L.; Bilger, C.; Lansing, C. R.; O'Brien, W. D. Jr.; Wheeler, B. C.; Feng, A. S. (2004). "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," J. Acoust. Soc. Am. 115 (1), January 2004, pp. 379-391.

Lockwood, M. L.; Jones, D. L. (2006). "Beamformer Performance With Acoustic Vector Sensor In Air", J. Acoust. Soc. Am, 119, 608-619, January 2006.

Ma, J.; Hu, Y.; Loizou, P. C. (2009). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions ", J. Acoust. Soc. Am., pp-3387-3405, May 2009.

Makinen, J.; Bessette, B.; Bruhn, S.; Ojala, P.; Salami, R.; Taleb, A. (2005). "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services", Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (ICASSP '05), Vol. 2, pp. 1109 -1112.

Manolakis, D. G.; Ingle, G. K.; Kogon, S. M. (2005)., "Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing" Boston: Artech House, INC.

Pulkki, V. (2007). "Spatial sound reproduction with directional audio coding," J. Audio Eng. Soc., vol. 55, no. 6, pp. 503–516, June 2007.

Schmidt, R. O. (1979).,"Multiple Emitter and Signal Parameter Estimation," Proceedings,RADC Spectral Estimation Workshop, 243-258, October 1979.

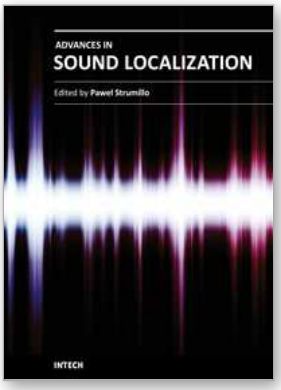
Shujau, M.; Ritz, C.H.; Burnett, I.S. (2009). "Designing Acoustic Vector Sensors for localization of sound sources in air", Proc. EUSIPCO 2009, Glasgow, Scotland , pp 849-853, 24-28 August 2009.

Shujau, M., Ritz, C.H., Burnett, I.S. (2010). "Source Separation using Acoustic Vector Sensors", Proc. IEEE 2010 International Conference on Acoustics, Speech and Signal Processing (ICASSP'2010), Dallas, Texas, March 14-19.

SoundField. User Manual for ST 250, Sound field reserch Ltd, West Yorkshire, England, Issue 1.5.

Ward, D. B.; Elko, G. W. (1999). "Robust and adaptive spatialized audio for desktop conferencing," Journal of the Acoustical Society of America, vol. 105, no. 2, p. 1099, Feb. 1999.

Wrigley, Stuart N.; Tucker, Simon; Brown, Guy J.; Whittaker, Steve (2009). "Audio spatialization strategies for multitasking during teleconferences", Proceedings of the Interspeech 2009, Pages 2935- 2938, September, 2009.



Advances in Sound Localization

Edited by Dr. Pawel Strumillo

ISBN 978-953-307-224-1
Hard cover, 590 pages
Publisher InTech
Published online 11, April, 2011
Published in print edition April, 2011

Sound source localization is an important research field that has attracted researchers' efforts from many technical and biomedical sciences. Sound source localization (SSL) is defined as the determination of the direction from a receiver, but also includes the distance from it. Because of the wave nature of sound propagation, phenomena such as refraction, diffraction, diffusion, reflection, reverberation and interference occur. The wide spectrum of sound frequencies that range from infrasounds through acoustic sounds to ultrasounds, also introduces difficulties, as different spectrum components have different penetration properties through the medium. Consequently, SSL is a complex computation problem and development of robust sound localization techniques calls for different approaches, including multisensor schemes, null-steering beamforming and time-difference arrival techniques. The book offers a rich source of valuable material on advances on SSL techniques and their applications that should appeal to researches representing diverse engineering and scientific disciplines.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Christian H. Ritz, Muawiyath Shujau, Xiguang Zheng, Bin Cheng, Eva Cheng and Ian S Burnett (2011). Backward Compatible Spatialized Teleconferencing based on Squeezed Recordings, Advances in Sound Localization, Dr. Pawel Strumillo (Ed.), ISBN: 978-953-307-224-1, InTech, Available from: http://www.intechopen.com/books/advances-in-sound-localization/backward-compatible-spatialized-teleconferencing-based-on-squeezed-recordings

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821