Speaker Diarisation of Vibroacoustic Intelligence from Drone Mounted Laser Doppler Vibrometers

Speaker Diarisation of Vibroacoustic Intelligence from Drone Mounted Laser Doppler Vibrometers

J L Richmond[1,*] & B J Halkon[1,2]

[1] School of Mechanical and Mechatronic Engineering, Faculty of Engineering and Information Technology, Ultimo, NSW 2007, Australia
[2] Centre for Audio, Acoustics and Vibration, Faculty of Engineering & IT, University of Technology Sydney, Ultimo, NSW 2007, Australia

[*]Josef.Richmond@student.uts.edu.au

Abstract. Laser Doppler Vibrometers (LDVs) are exceptionally well suited to non-contact vibration sensing applications in various environments. This work focuses on diarisation of conversations that might be recorded via a drone-mounted LDV by reducing the effect of external noise, extracting useful features from frames of audio and clustering them into homogenous segments based on speaker identity. The two-step noise reduction (TSNR) technique was introduced to these vibroacoustic data for the first time and tested against Gaussian bandpass filtering for noise reduction from sources such as laser speckle and additional broadband 'white' noise. Feature extraction was then performed using a time-delay neural network, with the grouping of frames to a particular speaker tested with various clustering methods. Each noise reduction and clustering technique combination were tested on a twospeaker conversation recorded via the LDV. In the case of no added noise, the most effective combination was found to be the TSNR/Agglomerative Hierarchical Clustering (AHC) combination with a diarisation error rate of 6.13%. In the case of additional broadband noise, the most effective combination was found to be TSNR followed by Gaussian bandpass filtering then clustering via AHC with a diarisation error rate of 11.9%. With this work, another aspect of the challenge of covertly obtaining and interpreting vibroacoustic intelligence in remote and hostile environments using LDVs has been addressed.

1. Introduction

Laser Doppler Vibrometry originates from the 1964 alternative to examining flow streamlines of injected dyes to measure fluid flow velocity [1]. This now well-established non-contact surface vibration measurement technique can be extended to the situation in which pressure waves produced by human speech cause nearby objects to vibrate. Using a laser Doppler vibrometer (LDV), the object's vibrations can be measured, and the original speech signal acquired [2],[3]. This speech acquisition can be performed remotely and unobtrusively by mounting the LDV to a drone [4]. The synthesis of these two technologies will potentially enable highly sensitive, non-invasive and discrete vibroacoustic intelligence-gathering activities in hostile environments without risk to human life. Once collected and processed, such intelligence can be used to create a transcript of the acquired conversation. An example of this process is shown in **Figure 1 (a)**. Furthermore, the actual experimental setup used to collect the data tested throughout this paper is described in **Figure 1 (b).**

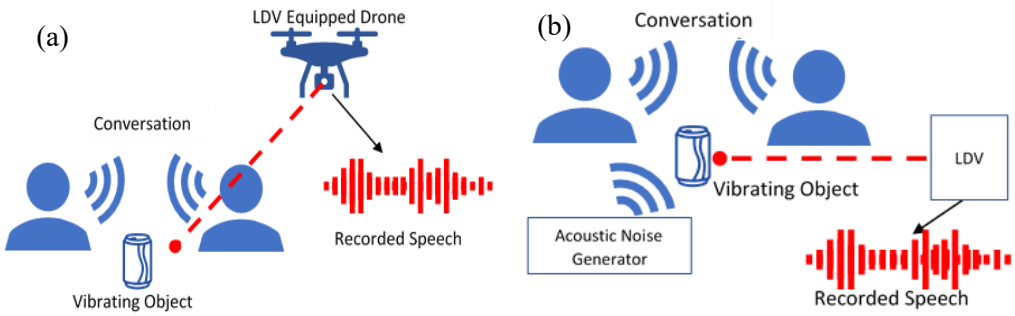

Figure 1 – (a) Example vibroacoustic intelligence-gathering schematic. An LDV-equipped drone measures the vibrations of a nearby drinks can caused by the pressure waves that propagate during a conversation. (b) Actual experimental setup using a Polytec NLV-2500-5 Compact Laser Vibrometer directed towards an aluminium can positioned nearby two male, English first language speakers, each speaking one sentence in turn from a pre-determined transcript.

Speaker diarisation – the process of partitioning an input audio stream into homogeneous segments according to the speaker identity – is a key part of the transcription process that has seldom been studied in the context of vibroacoustic data of the type considered here. In order to achieve effective diarisation of vibroacoustic intelligence acquired from a drone-mounted LDV, several engineering challenges must be overcome. Such an unconventional measurement scenario involves several sources of noise that are not easily controlled, quantified or characterised outside of the lab environment. These include but may not be limited to background acoustic noise, noise due to the laser speckle effect [1], shot noise [5] and noise from instrument vibration [6]. Furthermore, to determine the similarity between frames of speech and group these by speaker, features must be extracted from the audio signal that readily allows for discrimination between these different speakers. Finally, an appropriate clustering technique must then be utilised to group the speech frames. **Figure 2** shows the methodology proposed.

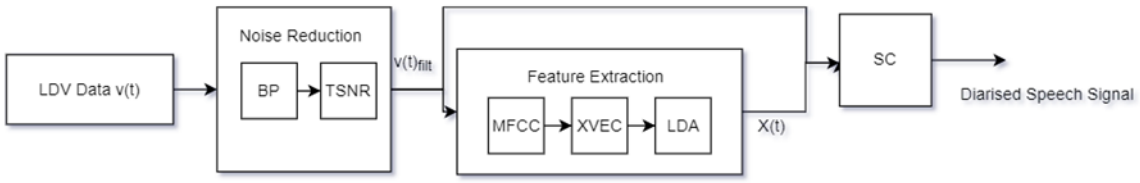

Figure 2 - Proposed speaker diarisation system. BP - Gaussian Bandpass Filter, TSNR – Two-Step Noise Reduction, MFCC – Mel Frequency Cepstral Coefficient Extraction, XVEC – X-Vector Feature Extraction, LDA – Linear Discriminant Analysis, SC – Segmentation and Clustering

2. Methodology

2.1. Noise Reduction

The use of LDVs for the measurement of speech signals in real-world scenarios, where a variety of noise sources are present, necessitates using noise reduction techniques to improve the quality of the speech signal reconstruction. However, the application of noise reduction techniques to vibroacoustic data of this specific nature has received relatively little attention in the published literature. Prior work has utilised Gaussian bandpass filters in combination with adaptive volume control to reduce signal power outside the human voice frequency range (300 – 3000 Hz) and enhance the speech signal [7]. This technique is limited as noise within the vocal frequency range is ignored. This work instead utilises the two-step noise reduction (TSNR) technique [8], a process used to estimate the *a priori* signal-to-noise ratio (SNR), thereby optimising a multiplicative gain function to filter out noise while avoiding the

addition of artificial high-power regions in the spectrum of a signal known as 'musical noise' in techniques using spectral filtering [9].

In this technique, the additive noise model is used, i.e. speech is modelled as the combination of a speech signal *s*(*t*) and a noise signal *n*(*t*), resulting in noisy speech *y*(*t*),

$$y(t) = s(t) + n(t) \tag{1}$$

Consider $Y(p, \omega_k), S(p, \omega_k)$ and $N(p, \omega_k)$ as the w_{Kth} spectral component of the time frame p of the noisy speech, speech, and noise signals, respectively. By applying a spectral gain $G(p, \omega_k)$ to each frame or sample of the noisy signal in frequency domain $Y(p, \omega_k)$, the additive noise can be partially filtered out. In general, this spectral gain is a function of both the *a priori* signal-to-noise ratio (SNR) and *a posteriori* SNR given respectively as,

$$\widehat{SNR}_{post}(p, \omega_k = \frac{|Y(p, \omega_k)|^2}{E[|N(p, \omega_k|^2]} \tag{2a}$$

$$\widehat{SNR}_{prio}(p, \omega_k) = \frac{E[|S(p, \omega_k)|^2]}{E[|N(p, \omega_k)|^2]} \tag{2b}$$

$E[]$ denotes expectation. Regardless of the type of filtering function used, spectral subtraction methods typically result in the addition of musical noise. In the TSNR process, the noisy signal is filtered through two Wiener filters. The effect of including the second gain function is to compensate for the musical noise effect introduced by the first. The first gain function is determined using the classical estimation of the *a priori* SNR,

$$G_1(p, \omega_K) = \frac{\widehat{SNR}_{prio}(p, \omega_k)}{1 + \widehat{SNR}_{prio}(p, \omega_k)} \tag{3}$$

while the second uses $G_1(p, \omega_k)$ to refine the *a priori* SNR for use in another Wiener filter,

$$\widehat{SNR}_{prio(2)}(p, \omega_k) = \frac{|G_1(p, \omega_K)\, Y(p, \omega_k)|^2}{\hat{\gamma}_N(p, \omega_k)} \tag{4}$$

$$G_2(p, \omega_K) = \frac{\widehat{SNR}_{prio(2)}(p, \omega_k)}{1 + \widehat{SNR}_{prio(2)}(p, \omega_k)} \tag{5}$$

Where $\hat{\gamma}_N(p, \omega_k)$ is the estimated noise power spectral density. The clean speech signal can then be estimated as,

$$\hat{S}(p, \omega_k) = G_2(p, \omega_k)\, Y(p, \omega_k) \tag{6}$$

By combining TSNR with a Gaussian bandpass filter to restrict the signals to the human vocal frequency range, the input signal can be cleaned of noise artefacts produced from various background vibroacoustic sources. The TSNR method was tested on samples of two different types: i) no extra noise, i.e. only inherent, typically low-level background noise, and ii) where broadband white noise was added using a simple smartphone app positioned near the speakers. The Gaussian bandpass filter was also tested both independently and in combination with the TSNR method. **Figure 3** shows the results of these various noise suppression options in both time and frequency domains. The original (A) time-domain signal quite clearly shows the two distinct speaker voices. While this distinction appears to be largely lost after bandpass filtering only (B), these differences are preserved and enhanced after TSNR processing only (C). There appears to be little to choose between the combined scenarios, (D) and (E),

in either time or frequency domain, with neither apparently being better than TSNR alone. A window length of 25 ms was used for the TSNR technique resulting in approximately 440 frames or samples from the 11 second recording.

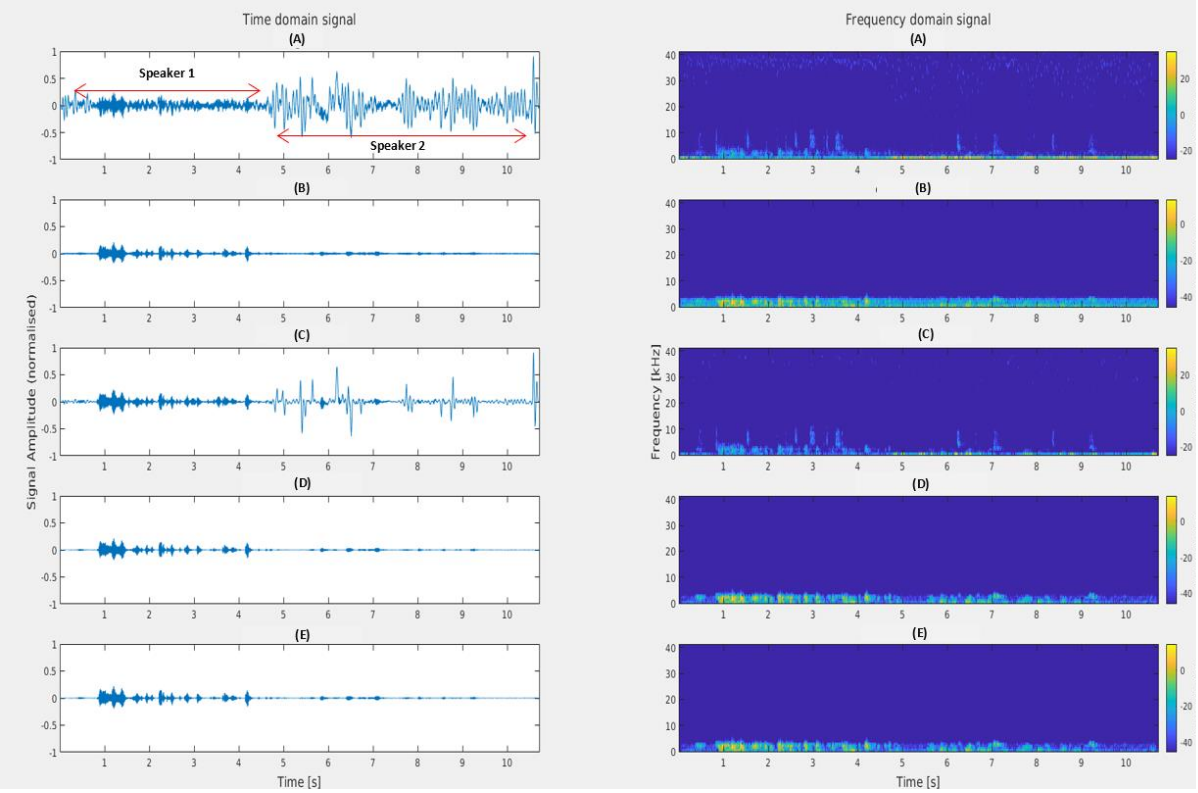

Figure 3 - Results of noise reduction techniques in time and frequency domain for signal with no added white noise. Row **(A)** shows the original signal with processing. Subsequent rows **(B-E)** show results of Gaussian bandpass filtering, TSNR filtering and combined filtering respectively. Portions of signal originating from each speaker roughly indicated in row **(A)**.

2.2. Feature Extraction

Given that speech is a complex information transfer, conveying multiple modalities highly dependent on tone, language, and the speaker, a representation of a frame of audio containing speech unique to a specific speaker, is needed. For example, many systems utilise Mel Frequency Cepstral Coefficients (MFCC) – coefficients that relate a short-term power spectrum to the human auditory system response – as features. The Mel frequency cepstrum, from which the MFCCs are extracted, differs from a typical cepstrum in that the frequency bands are based on the Mel scale, a perceptual scale of pitches designed to account for the nonlinear sensitivity of the human auditory system to different frequencies. However, the work presented by Kinnunen et al. [10] indicates that MFCCs do not cluster in feature space, suggesting that it can be expected that they are not an optimal way of performing speaker diarisation.

An alternate feature extraction method maps MFCCs to fixed-dimensional embeddings known as x-vectors using a time-delay neural network (TDNN) [11] with architecture as outlined in **Table 1**. The network was trained for speaker recognition using a downloaded database of approximately 2000 sentences spoken by ten male and ten female speakers captured via microphone [13]. With the network trained, the first six layers extract features from the neighbouring audio frames that readily discriminate between different voices. These features are pooled in layer 7 and extracted from the network to be subsequently used for clustering, which is subsequently described in section 2.3.

Table 1. TDNN architecture for classifying N speakers for T framed audio input. Layer context refers to the range of speech frames operated over by a layer centred at the current frame t, i.e. each frame is operated on with the neighbouring 4 frames from t – 2 to t + 2 in the first layer, giving a group size or total context of 5 frames. TDNN architecture is further discussed in [9].

Layer	Layer Context	Total Context	Input x Output
1	{t - 2, t + 2}	5	120 x 512
2	{t – 2, t, t + 2}	9	1536 x 512
3	{t – 3, t, t + 3}	15	1536 x 512
4	{t}	15	512 x 512
5	{t}	15	512 x 1500
6	[0, T)	T	1500 T x 3000
7	{0}	T	3000 x 512
8[a]	{0}	T	512 x 512
9[a]	{0}	T	512 x N

[a] Layer not used in the final feature extraction process.

Linear discriminant analysis (LDA) [13] was then applied to reduce the dimensionality of the extracted x-vectors, ideally allowing more efficient clustering while still conveying the information necessary to distinguish between the different speakers. For an approximately 11-second recording with a sampling frequency of 81.92 kHz, 2135, 30-dimensional MFCC features were extracted using a hop distance (time between centre of neighbouring frames) of 5 ms and a frame duration of 40. Using the trained neural network, 525, 120-dimensional x-vectors were then extracted from these MFCCs using a hop distance of 0.1 s and a frame duration of 2 s. Each of these x-vectors then had its dimensionality reduced to 32 dimensions through LDA. The plots in **Figure 4** show each stage of this feature extraction process, which converts the time domain signal recorded by the LDV into a set of feature vectors which should enable segmentation of homogenous regions according to the specific speaker.

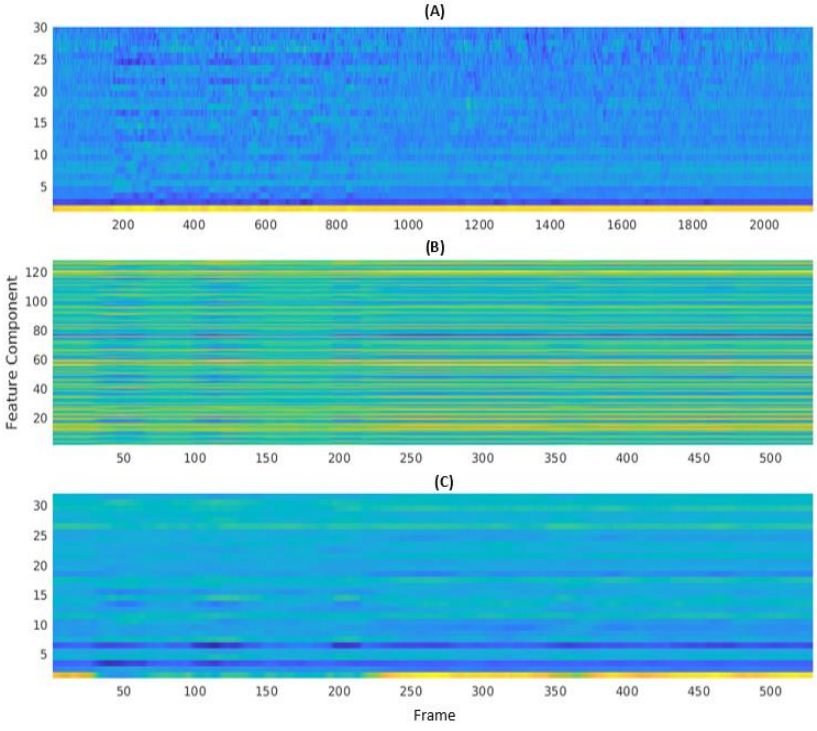

Figure 4 – Feature extraction from an input set of MFCCs. **(A)** – Normalised MFCCs extracted from the original signal in the frequency domain. **(B)** – X-vectors extracted from neighbouring groups of MFCCs. Number of segments reduced due to use of multiple neighbouring MFCC frames to extract a single X-vector. **(C)** - LDA dimensionality reduced x-vectors.

2.3. Clustering

The speaker diarisation process consists of two phases: segmentation and clustering. The former involves determining when a speaker change has occurred in the signal, while the latter involves grouping segments corresponding to the same speaker. Segmentation is addressed in this work by splitting the input audio into windowed and overlapped segments. Once extracted and reduced, the x-vectors for each frame can then be clustered by various methods such as k-means, agglomerative hierarchical clustering (AHC) and Gaussian mixture modelling [12]. Each of the three clustering methods was tested in combination with each of the five processed signals described in section 2.1, resulting in a total of 15 test cases for each of the no added noise cases and the case where broadband noise was added. **Figure 5** presents the results of a diarisation on the no added noise signal using the TSNR filtering and AHC processes, i.e., the features extracted from windowed segments of the original signal have been clustered into three groups, with group zero corresponding to frames in which neither person was speaking.

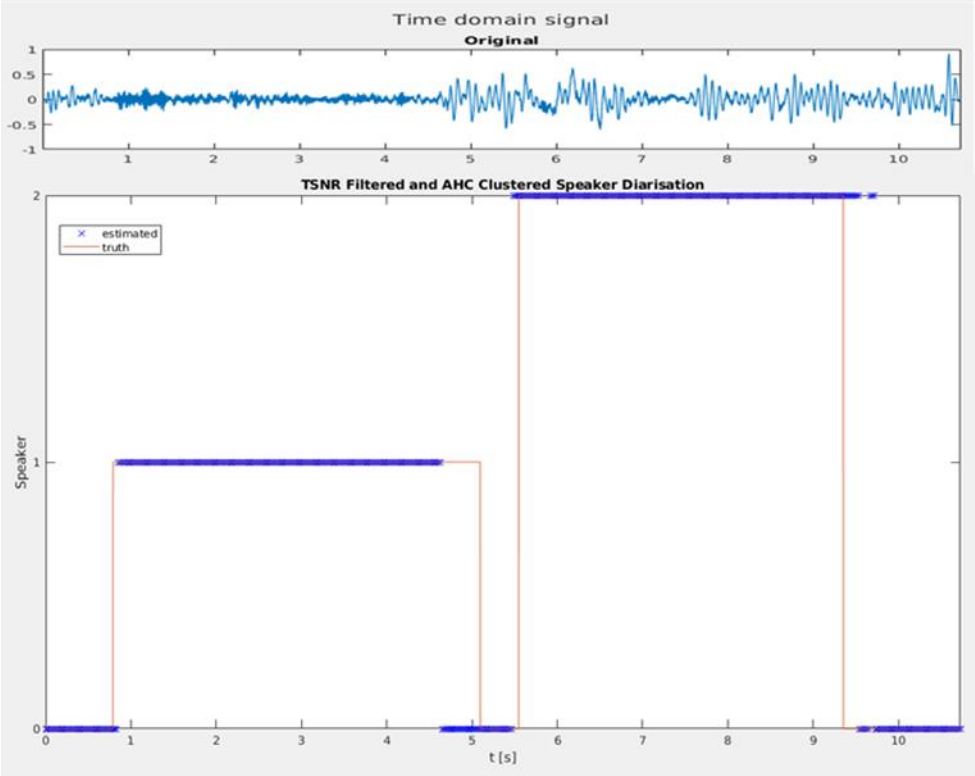

Figure 5 - Diarisation results of the original signal with no added noise filtered by TSNR and clustered via AHC. The orange line indicates the approximate ground truth, with values of zero indicating silence. Blue crosses indicate the results of the clustering process, i.e. the estimation of who is speaking in each frame.

3. Results and Discussion

In order to assess the effectiveness of the various noise reduction / diarisation process combinations, the diarisation error d, calculated according to the number of frames assigned to the incorrect cluster e and the total number of frames N, was defined,

$$d = \frac{e}{N} \tag{7}$$

The diarisation errors for each process combination in both the no noise and added white noise cases are presented in **Table 2**.

Table 2 - Noise reduction / clustering diarisation error, *d*. The most effective methods for each noise case are shaded and shown in **bold,** while the least effective are shown in *italics*.

Noise Treatment Method	No added noise			Added white noise		
	k-means	AHC	GMM	k-means	AHC	GMM
No Noise Reduction (1)	0.102	0.118	0.105	0.340	*0.347*	0.301
Gaussian Filtering (2)	*0.131*	0.100	0.187	0.295	0.275	0.281
TSNR (3)	0.0642	**0.0613**	0.0737	0.223	0.149	0.244
Gaussian then TSNR (4)	0.112	0.100	0.113	0.293	0.286	0.239
TSNR then Gaussian (5)	0.104	0.0945	0.134	0.128	**0.119**	0.155

In the case of no added noise, the combination of TSNR only and AHC resulted in the most effective diarisation with an error rate of 6.13%. This represents an error rate of approximately half that of the equivalent case without noise reduction treatment while still using AHC for clustering. The worst performing combination for the no added noise scenario was Gaussian filtering only combined with k-means clustering; an error rate of 13.1% was observed. Irrespective of the clustering technique, preceding with TSNR only noise reduction offers the best performance. It is also noteworthy that preceding with the Gaussian bandpass filter only resulted in the worst combined performance for all clustering approaches; in two of the three cases, the effectiveness was in fact reduced when compared to the corresponding case where no noise reduction technique at all was used.

In the case where additional white noise was played while the speakers conversed, the most effective combination was TSNR then Gaussian bandpass filtering for noise reduction followed by AHC for clustering. Here, an error rate of 11.9% was observed with the inclusion of the Gaussian filtering after the TSNR offering an additional 3% performance increase. Compared with the corresponding combination for the no added noise case, this TSNR then Gaussian followed by AHC exhibited only a ~2.5% reduction in performance. The worst performing combination was no noise reduction followed by AHC with a substantial error rate of 34.7%. Irrespective of the clustering approaches, in the added noise case, a performance benefit was always observed by preceding with some form of noise reduction.

4. Conclusions

This paper has presented the development and testing of various methods for speech diarisation of vibroacoustic data intended for the end goal of enabling transcription and enabling drone-mounted laser Doppler vibrometer (LDV) for use in remote, non-invasive and covert intelligence gathering. The means by which this was achieved relied primarily on established signal processing techniques that have been applied in this article to vibroacoustic data for the first time. To test the performances of the various combinations of diarisation methods, new experimental were generated with an LDV measuring the surface vibration from an aluminium drink can in the proximity of two male speakers. Reciting a pre-determined transcript and talking in turn, two cases were conceived with and without the deliberate addition of broadband white noise from a simple smartphone app.

As an initial step, two methods for compensating for the noise in the LDV measurement were investigated. A Two-Step Noise Reduction (TSNR) technique was and was shown to outperform the Gaussian filtering technique by inspection of time and frequency domain data. Subsequently, a trained deep neural network-based feature extraction technique was utilised to produce features for speech frames that readily correlate to the different speakers and overcome the non-clustering behaviour of Mel Frequency Cepstral Coefficients (MFCC). Several clustering techniques were then tested on the extracted features in order to diarise the speech. This process was used for both the no noise case and the added white noise case, with each combination of noise reduction/clustering techniques being tested. In the case where no noise was added, the most effective combination was TSNR/AHC with a diarisation error rate of 6.13%, approximately half the error rate of the corresponding case where no noise suppression was used. For the added white noise case, the combination of TSNR then Gaussian bandpass filtering followed by AHC clustering was the most effective with an error rate of 11.9%, approximately a third of the error rate of the corresponding no noise suppression case. This technique combination also

transferred particularly well compared to the others, with only a ~2.5% decrease in performance compared to the no noise case.

Future research should investigate improvements in performance possible through training of the deep neural network with noisy speech signals, perhaps those recorded with LDVs in place of the freely available 2000 sentences for male and female speakers used in this study. Furthermore, the potential for live speaker diarisation of LDV data via techniques such as real-time exponential filter clustering to enable real-time noise reduction and transcription of vibroacoustic intelligence should be completed. This will also necessitate optimising noise reduction and feature extraction techniques to reduce lag time as much as possible. A cost-benefit analysis of each of these various techniques should be conducted to better optimise for rapid, in-field implementation of speech diarisation. With the work presented here, another aspect of the challenge of covertly obtaining and interpreting vibroacoustic intelligence in remote and hostile environments has been investigated.

References

[1] Rothberg S J, *et al*. 2017 An international review of laser Doppler vibrometry: making light work of vibration measurement *Optics and Lasers in Eng.* **99** 11-22.
[2] Xiao T, Zhao S, Qiu X and Halkon B J 2021 Using a retro-reflective membrane and laser Doppler vibrometer for real-time remote acoustic sensing and control *Sensors*
[3] Xiao T, Qiu X and Halkon B J 2020 Ultra-broadband local active noise control with remote acoustic sensing *Sci. Rep.* 20784(2020)
[4] Halkon B J, Rothberg S J 2018 Towards laser Doppler vibrometry from unmanned aerial vehicles *Journal of Phys: Conf Series* p 1149
[5] Rembe C, Kowarsch R 2017 High resolution laser-vibrometer microscopy *Proc. Sensor 2017 (Nürnberg)* p 244
[6] Halkon B J and Rothberg S 2017 Taking laser Doppler vibrometry off the tripod: correction of measurements affected by instrument vibration *Optics and Lasers in Engineering* **91** 16-23
[7] Zhu Z, Li W and Wolberg G 2005 Integration of laser vibrometry with infrared video for multimedia surveillance display *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (San Diego)* p 478
[8] Plapous C, Marro C, Mauury L and Scalart P 2004 A two-step noise reduction technique *Proc. IEEE Int. Conf.: Acoustics, Speech and Sig. Proc (Philidelphia)* p 289
[9] Chinaev A and Hab-Umbach I 2016 A priori SNR estimation using a generalised decision directed approach *Proc. Interspeech (San Diego)* p 474
[10] Kinnunen T, Karkkainen I and Franti P 2001 Is speech data clustered? – statistical analysis of cepstral features *Proc. Interspeech (Aalborg)* p 2627
[11] Snyder D, Garcia-Romero D, Sell G, Povey D & Khudanpur S 2018 X-vectors : robust DNN embeddings for speaker recognition *Proc. Int. Conf. on Acoustics, Speech and Signal Proc. (Calgary)* p 5329
[12] Pernkopf F 2011 PTDB-TUG : Pitch tracking database from Graz university of technology *Graz University of Technology*
[13] Izenman A J 2013 *Linear discriminant analysis* in Modern multivariate statistical techniques (New York, NY Springer) pp 237-280