Clustered Multi-channel Dereverberation for Ad-hoc Microphone Arrays

Shahab Pasha and Christian Ritz

School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, Australia

***Abstract*—** A novel unsupervised multi-channel dereverberation approach in ad-hoc microphone arrays context based on removing microphones with relatively higher level of reverberation from the array and applying the dereverberation method on a subset of microphones with lower level of reverberation is proposed in this paper. This approach does not require any prior information about the number of microphones and their relative locations, however based on kurtosis of Linear Prediction (LP) residual signals, microphones located close to the active source are detected and utilized for the dereverberation process. The proposed method is a clustered enhancement method which can be applied with any dereverberation algorithm. The proposed method is not dependent on the recording setup so it requires no predefined threshold and it can be applied to unknown rooms with unseen speakers. Dereverberation results suggest that regardless of the applied dereverberation method, using a consciously chosen subset of microphones always yield better dereverberation results compared to blind use of all microphones.

I. INTRODUCTION

In recent years ad-hoc microphone arrays, which are formed from randomly distributed microphones, have been widely used for recording and analyzing acoustic scenes within a large space such as a room due to their spatial coverage and flexibility (e.g. by forming arrays from microphones attached to mobile devices) [1,2]. Although compared to fixed geometry arrays such as the Uniform Linear Array (ULA), ad-hoc microphone arrays are more effective tools for recording and analyzing acoustic scenes [3], speech enhancement in this context is still challenging and complicated as despite compact arrays, there is no information about the relative distances and time delays between the channels. Moreover each single microphone in an ad-hoc array has its own unique and distinctly different Room Impulse Response (RIR) and echo pattern which means sound reflections are not consistent within the array. It is shown in [4,5] that it is possible to suppress reverberation and cancel the effect of echoes if the microphone array geometry (i.e. time delays) is known but these methods are not directly applicable to a general scenario of randomly distributed microphone array.

In a recent research [6] a novel speech enhancement by randomly distributed compact microphone arrays is introduced and tested. The norms of the pseudo-coherence vectors and Signal to Noise Ratio (SNR) within each compact array are utilized as array selection criteria. In other words selection criteria as mentioned above are applied to choose a subset of compact arrays that yield better speech enhancement results. It is concluded that the proposed criteria are effective selection features to choose a subset of arrays prior to the beamfoming phase.

Although clustered based approaches to speech enhancement and speaker activity detection techniques with ad-hoc known geometry microphone arrays [6,7] are shown to be efficient and effective tools, speech enhancement in ad-hoc single microphone arrays (where each node consists of only one single microphone and not a compact array), which is a more common scenario in applications such as meetings and interviews is not studied and investigated enough and most of the criteria suggested for fixed geometry microphone array processing (e.g. output SNR and intra node coherence) are not applicable to ad-hoc single microphone arrays.

In [8] the authors have utilized the observation that Linear Prediction (LP) residual signals of clean (not reverberant) speech signals have strong, distinct peaks that corresponds to pulses generated by the vocal cords but reverberant speech signals have spreading random peaks over time. This difference between clean and reverberant signals is utilized to discriminate close and distant speakers by one single microphone. In other words, that research uses reverberation to compare and discriminate sources (speakers) based on their relative distances to the microphone without any prior knowledge of microphone and sources relative positions. The average Kurtosis of LP residual signals over a number of frames from an active speaker is compared with a predefined threshold to distinguish close and distant speakers. The authors suggest that determining a suitable threshold should be investigated more in the future. Moreover defining the threshold needs training which is highly dependent on the acoustic environment characterized by the wall absorption factors, reverberation time (RT_{60}) and speaker positions, therefore using a threshold to discriminate close and distant speakers in a supervised manner cannot be generalize to all setups which is a limitation to that approach.

Based on the proposed close/distant talker discriminative feature in [8] herein a novel clustered dereverberation method in ad-hoc single microphone array context is proposed. As the proposed method is an unsupervised clustering method it overcomes the drawback of [8] which is the need of a predefined or trained threshold. Moreover the proposed method is applicable to ad-hoc single microphone arrays where time delays between microphones are not known (limitation of [4,6]). The proposed microphone discriminative feature in this paper can be applied to single microphones as well, so it is a more general feature compared to the proposed criteria in [6] which is only applicable to compact microphone array nodes. As the proposed feature in this paper is a relative value the proposed method is robust against RIR changes and despite the applied approach in [6] there is no need to assume that RIRs are fixed during the experiments.

The remainder of the paper is organized as follows. Section II is dedicated to problem formulation in a general scenario. Section III describes a discriminative feature for microphone clustering based on reverberation level and shortly explains the applied state of the art dereverberation methods. Experiments and results are represented in Section IV. In section V the paper is concluded.

Ad-hoc microphone array recording in a reverberant environment When speech signal s(n) is recorded in a noisy reverberant room, its quality is downgraded by reverberation and noise. Reverberation is more challenging because it has a long term effect that distorts

Acquire all M channels signals: $\boldsymbol{x_1(n)} \dots \boldsymbol{x_m(n)} \dots \boldsymbol{x_M(n)}$
Obtain the p order LPC coefficients for each channel separately $\begin{bmatrix} \boldsymbol{a_{11}} & \cdots & \boldsymbol{a_{1p}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{a_{M1}} & \cdots & \boldsymbol{a_{Mp}} \end{bmatrix}$
Reconstruct the signal based on LPC coefficients for all channels. $\hat{\boldsymbol{x}}_1(\boldsymbol{n}) \dots \hat{\boldsymbol{x}}_m(\boldsymbol{n}) \dots \hat{\boldsymbol{x}}_M(\boldsymbol{n})$
Obtain residual signal for all channels: $\boldsymbol{e_1(n)} \dots \boldsymbol{e_m(n)} \dots \boldsymbol{e_M(n)}$
Calculated the discriminative feature $\boldsymbol{k_1(n)} \dots \boldsymbol{k_m(n)} \dots \boldsymbol{k_M(n)}$
Apply the unsupervised clustering algorithm on discriminative features values
Apply the dereverberation method on the formed clusters

Fig.1: Proposed method

several time frames, this issue can cause more distortion if reverberation time (RT_{60}) is large (i.e. beyond 1s).

Although the recording setup is equal for all single microphones in ad-hoc arrays the quality of signals recorded by microphones located far from the source is downgraded more than other microphones. The goal of this research is to remove microphones highly affected by reverberation from the array and apply the dereverberation methods (i.e. delay and sum beamforming and Multi-channel LPC) only on microphones with lower levels of distortion in order to achieve a more effective dereverberation tool.

Reverberation can be modeled by convolving the clean signal with the L-tap RIR at each microphone position $h_m = [h_{m,0}, h_{m,1}, \dots, h_{m,L}]$ where L is the number of significant echoes and m is the microphone index. Recorded distorted signals by each single microphone consists of three parts: a) direct path clean signal, b) Echoes and reflections and c) Noise

$$x_m(n) = x_{m,E}(n) + x_{m,L}(n) + v(n) \quad (1)$$

$$x_m(n) = h_m^T * s(n) + v(n), \quad m = 1,2,..,M \quad (2)$$

Where $v(n)$ is the noise signal recorded by m^{th}channel, * denotes the convolution operator and M is the total number of single microphones in the ad-hoc microphone array. Although reverberation is usually considered as a source of distortion, it can contain helpful information. In [3] up to 15^{th} order reflections (i.e. $L=15$) have been applied for source localization by compact and ad-hoc single microphone arrays and it is shown that due to their flexible and wide spatial coverage, ad-hoc microphone arrays can analyze an acoustic scene (e.g. source localization) more accurate than compact arrays.

In this research in a general scenario of M single (not collocated) microphones, randomly distributed in a reverberant room at unknown positions, the objective is to choose a subset of microphones such that applying the dereverberation process leads to the highest level of improvement in speech quality and echo cancellation. The hypothesis of this research is that excluding highly reverberated microphone signals from the dereverberation process can improve the results.

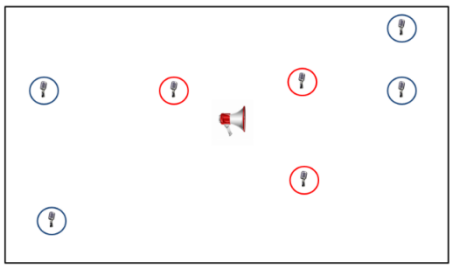

Fig. 2: Clustering based on kurtosis of LP residuals

II. CLUSTERED DEREVERBERATION

In some recent research it is shown that reverberation can be exploited to obtain information about the recording environment such as room geometry and source locations [8,9,10,11]. Inspired by those works and given that clustered and informed approaches are shown to yield better results in ad-hoc microphone array contexts [1,2,6,12], this research is trying to find a setup independent approach to choose a subset of microphones that yields higher quality outputs in terms of echo cancellation. In order to achieve this goal the first step is to extract discriminative features from speech signals to choose a subset of single microphones (Section III.A).

In this research, the level of reverberation within each channel is estimated and applied as an indicator to distinguish microphones with relatively high/low signal qualities. Delay and sum beamforming and multi-channel LPC are applied then on clustered microphones to suppress the reverberation. The applied machine learning technique is an unsupervised method however based on the analysis of received signals an informed dereverberation process is introduced (Section III.B). As blind approaches in ad-hoc microphone arrays context need to overcome the problems of microphone and source localizations, channel synchronization and gain equalization [13], in this paper an informed, setup independent approach without prior information is implemented and tested (Section IV).

The dereverberation process contains two phases, phase one is choosing a subset of microphones that yields a higher output quality compared to blind use of all microphones in the array and phase two is applying a multi-channel dereverberation approach on the chosen subset (Fig.2). As a general scenario consider a randomly distributed microphone array of M single microphones at unknown locations and one active source at an unknown position.

Recorded speech signal by the m^{th} channel is represented as $x_m(n)$ which is sampled by sampling rate f_s. LP coefficients derived from the recorded signal can represent the signal as a function of p previous samples:

$$\hat{x}_m(n) = f(x_m(n-1), \dots, x_m(n-p)) \quad (3)$$

Where p is the order of LP analysis. LP coefficients are then utilized to calculate the estimated signal $\hat{x}(n)$ based on P previous samples.

$$\hat{x}_m(n) = -\sum_{i=1}^{p} a_i \, x_m(n-i) \quad (4)$$

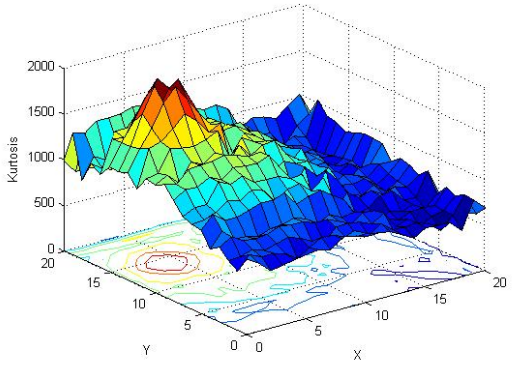

Fig. 3: Scaled kurtosis of LP residual signals for 400 microphones evenly distributed at the height of z=2m

The residual signal $e_m(n)$ can be obtained by calculating the difference between the original recorded signal $x_m(n)$ and the reconstructed estimated signal $\hat{x}_m(n)$ by (4).

$$e_m(n) = x_m(n) - \hat{x}_m(n) \tag{5}$$

The discriminative feature used in this research is the kurtosis of the residual signal $e_m(n)$ of each channel [8] which is obtained by calculating the kurtosis of (5):

$$k_m(n) = \frac{E\{e_m^4(n)\}}{E^2\{e_m^2(n)\}} - 3 \tag{6}$$

Where E{.} denotes the mathematical expectation operator. As suggested in [8], a frame based process is applied for calculating the kurtosis of LP residual signals. The average kurtosis of *N* short frames (i.e. 20ms) calculated by (6) can be applied as the discriminative feature within machine learning methods as:

$$\bar{k} = \frac{1}{N}\sum_{l=1}^{N}\frac{K(l)}{L} \tag{7}$$

A. Kurtosis of LP residual as a discriminative feature

LP residuals of clean signals contain distinct peaks at Glottal Closure Instants (GCI) and very low values between these peaks whereas reverberated signals are not following the speakers pitch exactly as original peaks are distorted and artificial peaks are generated by reverberation. This difference between clean and reverberated signals is utilized in [8] to discriminate close and distant talkers but defining close and distant, requires a threshold which highly relies on each specific setup. In order to avoid this limitation, in this research kurtosis of LP residual signal is used within unsupervised machine learning algorithms.

As it is demonstrated in Fig. 3 in a 20m by 20m by 3m room and a source positioned at 5m,15m and 2m, the peak of the kurtosis of LP residual signal graph obtained by (6,7) on a 2D plane grid at fixed height of 2m with 1m step size, is around the source position and it decays with distance from the source. In other words kurtosis of LP residual signal has an inverse relationship with distance to the source. This observation inspires using kurtosis of LP residual signals as a reliable discriminative feature to discriminate relatively far and consequently highly reverberated microphones from relatively close and cleaner microphones signals in an unsupervised manner without any predefined setup dependent threshold.

B. Unsupervised microphone discrimination and clustering

In order to determine if a channel is far (highly reverberated) or not, a reverberation threshold needs to be defined, in [8] trial and error approach (0 to 20 with 0.01 step size) is applied to choose a suitable threshold for kurtosis values and authors suggest more work is needed to be done on this part. Apart from the problem of choosing a threshold value, the threshold is not independent from the setup and it needs to be updated for each recording environment. As the optimized threshold is always defined with uncertainty, kurtosis of LP residual signal as calculated by (6,7) can be applied as a discriminative feature for clustering microphones into two clusters without any predefined threshold in an unsupervised manner. The number of clusters is a critical issue in all unsupervised clustering methods, in this research as the goal is to decide if a microphone is highly reverberated or not (located far from the source or not), there are always two (K=2) non-empty clusters (cluster far and cluster close). Standard K-means clustering as explained in [14] is implemented and applied to microphone clustering based on their kurtosis of LP residual signals. Having microphones clustered into two clusters state of the art dereverbeartion can be applied on the clustered microphones.

C. Delay and Sum Beamforming (DSB)

State of the art approaches to multi-channel dereverberation try to attenuate the residual signals between GCIs as they are not generated by the speaker and they contain reverberation and echoes. In [4] DSB and the Spatiotemporal averaging of Method for Enhancement Reverberant Speech (SMERSH) are applied to suppress the reverberation between GCIs by compact microphone arrays. In this research the same approach is applied to distributed ad-hoc single microphones with required modifications.

For spatiotemporal averaging, delays between channels are required to time align the channels. Once the time aligned signals are obtained it is possible to suppress the uncorrelated parts by averaging [4]. In this research by calculating the cross-correlation between each channel and a reference channel (which can be chosen randomly) the relative delays between channels are obtained and utilized to time align the signals and average them to obtain the dereverberated signal $\bar{x}_{derev}(n)$as represented in (8):

$$\bar{x}_{derev}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n - d_m) \tag{8}$$

Where d_m is the delay between the reference and the m^{th} channel. Applying this process blindly to all microphones may not be the optimized approach in terms of calculation cost and the output quality. Here the process of dereverberation is applied to the subset chosen by the K-means clustering method. DSB results for the chosen subset (cluster close) and all the microphones in the array can be calculated
by (9,10).

$$\bar{x}_{Close}(n) = \frac{1}{M_{Close}}\sum_{m=1}^{M_{Close}} x_m(n - d_{m,ref})$$

$$\bar{x}_{All}(n) = \frac{1}{M_{All}}\sum_{m=1}^{M_{All}} x_m(n - d_{m.ref}) \tag{10}$$

Table 1: Experimental setup

Source signals	IEEE_Corpus wideband
Noise	White noise, 20 dB
LPC order	10
Frames size	20ms
Room dimensions	6m×5m×3m
Reverberation time (RT_{60})	200ms, 400ms

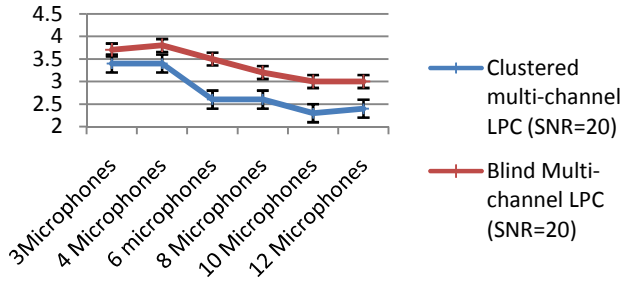

Fig. 4: Cepstral distance vs. total number of microphones

D. *Multi-channel LPC*

Using (4) a speech signal $x(n)$ can be represented by LPC coefficients and similarly all M reverberant signals recorded by M microphones can be written as $x_m(n) = \hat{x}_m(n) + e_m(n)$. It is shown that LP coefficients obtained by joint analysis of M reverberant channels can estimate the clean LP coefficients accurately however reverberation still exist in the residuals [4]. It is suggested that averaging time-aligned residual signals can suppress the uncorrelated part (i.e. reverberation) [4,16]. In [5] AutoRegressive (AR) models (e.g. LPC) are obtained from clean and reverberated signals and it is shown that the spatially expected values of the reverberant speech AR coefficients are approximately equal to those achieved by the clean signal. In other words if AR coefficients are derived from each reverberant channel separately (which is possible in an ad-hoc microphone array context) they converge to or cluster around the clean signal coefficients. In this research Line Spectral Frequency (LSF) coefficients derived from LPC coefficients are utilized for the averaging process as despite LPC coefficients, LSF coefficients are always positive and cancelling issue can be avoided. Method of [4] is applied to time align and average the residual signals. Having dereverberated LPC coefficients and the averaged residuals, the dereverberated signal can be achieved by (4,5).

III. EXPERIMENTAL SETUP AND REULTS

In a noisy, reverberant 6m×5m×3m room with one active source, 30 different setups of 3 to 12 microphones and a speech source at 4 different positions have been simulated. The dereverberation performance of blind use of all microphones is compared with the performance of the chosen subset (cluster). The chosen subset consists of microphones clustered as close by K-means (K=2) method due to their higher kurtosis of LP residuals. The source has been located at a range of different positions including at the center and very close to the reflectors (i.e. walls). Microphones are distributed in a wide range of distances from the source from 10 cm to 7m.

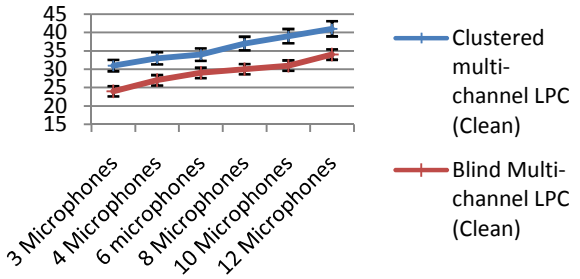

Fig. 5: DRR vs. total number of microphones

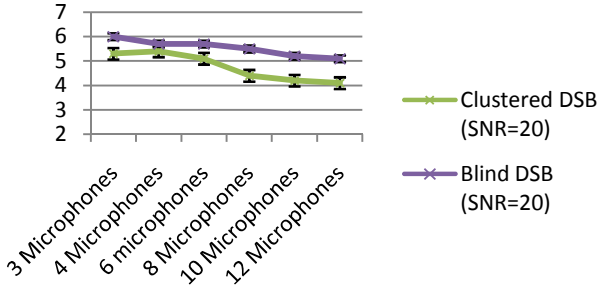

Fig. 6: Cepstral distance vs. total number of microphones

Two different reverberation times and noise levels have been applied to investigate the robustness of the results against environmental conditions.

In Fig.4 average cepstral distances for all setups in terms of source and microphone positions and reverberation time is represented for both clustered and blind approaches. It can be observed that regardless of the number of microphones clustered multi-channel LPC approach always yields better (Lower) cepstral distances between the clean source signal and the dereverberated output of the array. In Fig.5 Direct to Reverberation Ratio (DRR) is calculated as the dereverberation measurement and it is shown that applying the multi-channel LPC on a chosen subset of microphones, clearly yields better (higher) DRRs. Comparison of Multi-channel LPC and DSB is not an objective of this research but it is clearly shown that multi-channel LPC has a superior performance (Fig.4 and Fig 6).

IV. CONCLUSION

A novel unsupervised clustered dereverberation method utilizing kurtosis of LP residual signals as discriminative feature has been introduced and tested. The proposed method informs the dereverberation method of the microphones distances from the source and excludes highly reverberated signals from the dereverberation process. Multi-channel LPC and DSB have been implemented as state of the art reverberation suppression approaches in different setups in terms of the number of microphones, noise level and relative distances between microphones and the source. Results suggest that the proposed informed approach can always yield better results compared with the blind approach where all microphone are included. It can also be concluded that kurtosis of LP residual signal is a noise robust, setup independent and effective criteria for dereverberation applications in ad-hoc microphone arrays context.

REFERENCES

[1] Gergen, S., Nagathil, A, Martin, R., "Audio signal classification in reverberant environments based on fuzzy-clustered ad-hoc microphone arrays," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.3692,3696, 26-31 May 2013

[2] Himawan, I, McCowan, I, Sridharan, S., "Clustering of ad-hoc microphone arrays for robust blind beamforming," *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* , vol., no., pp.2814,2817, 14-19 March 2010

[3] Asaei, A., Bourlard, H., Taghizadeh, M.J., Cevher, V., "Model-based sparse component analysis for reverberant speech localization," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* , vol., no., pp.1439,1443, 4-9 May 2014

[4] Gaubitch, N.D., Naylor, P.A., "Spatiotemporal Averagingmethod for Enhancement of Reverberant Speech," *Digital Signal Processing, 2007 15th International Conference on* , vol., no., pp.607,610, 1-4July 2007 doi: 10.1109/ICDSP.2007.4288655

[5] N. D. Gaubitch, D. B. Ward and P. A. Naylor, "Statistical analysis of the AR modeling of reverberant speech". *J. Acoust. Soc. Am.,* vol. 120, no. 6, pp. 4031-4039, 2006.

[6] Tavakoli, V. , Jensen, J. Christensen, M and Benesty, J." Pseudo-Coherence-based MVDR beamforming for speech enhancement with ad-hoc microphone arrays" *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*

[7] Keisuke, K., Tomohiro N. "Audio and Acoustic Signal Processing: Audio and Speech Source Separation" *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*

[8] Hayashida, K., Nakayama, M., Nishiura, T., Yamashita, Y., Horiuchi, T., Kato, T., "Close/distant talker discrimination based on kurtosis of linear prediction residual signals," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* , vol., no., pp.2327,2331, 4-9 May 2014

[9] Takashima, R., Takiguchi, T., Ariki, Y., "Prediction of unlearned position based on local regression for single-channel talker localization using acoustic transfer function," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.4295,4299, 26-31 May 2013

[10] Dokmanic, I.,Parhizkar R., Walther A., Yue M., Vetterli M. "Acoustic echoes reveal room shape" *Proceedings of the National Academy of Sciences* 110, no. 30: 12186-12191. 2013

[11] Longbiao Wang, Zhaofeng Zhang; Kai, A., Kishi, Y., "Distant-talking speaker identification using a reverberation model with various artificial room impulse responses," *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific* , vol., no., pp.1,4, 3-6 Dec. 2012

[12] Vincent, E., Bertin, N., Gribonval, R., Bimbot, F., "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound," *Signal Processing Magazine, IEEE* , vol.31, no.3, pp.107,115, May 2014

[13] Gaubitch, N.D., Martinez, J., Kleijn, W.B., Heusdens, R., "On near-field beamforming with smartphone-based ad-hoc microphone arrays," *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on* , vol., no., pp.94,98, 8-11 Sept. 2014

[14] Rogers S., Girolami M., " A First Course in Machine Learning ", Chapman & Hall/Crc, October 2011

[15] Gillespie, Bradford W., Malvar, H.S., Florencio, D.A.F., "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on* , vol.6, no., pp.3701,3704 vol.6, 2001

[16] Shujau, M., Ritz, C.H., Burnett, I.S., "Speech dereverberation based on Linear Prediction: An Acoustic Vector Sensor approach," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.639,643, 26-31 May 2013