Downloaded from the open archive of Mid Sweden University DiVA.

This is the accepted version of a paper published in 2019 30th Irish Signals and Systems Conference (ISSC). See full citation below.

Pasha, Shahab; Lundgren, Jan, "Multi-Channel Compression and Coding of Reverberant Ad-Hoc Recordings Through Spatial Autoregressive Modelling," *2019 30th Irish Signals and Systems Conference (ISSC)*, 17-19 June 2019
<http://dx.doi.org/10.1109/ISSC.2019.8904918>

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

MULTI-CHANNEL COMPRESSION AND CODING OF REVERBERANT AD-HOC RECORDINGS THROUGH SPATIAL AUTOREGRESSIVE MODELLING

Shahab Pasha, Jan Lundgren

STC Research Centre, Mid Sweden University, Sweden

ABSTRACT

Autoregressive modelling techniques such as multi-channel linear prediction are widely used for applications such as coding, dereverberation and compression of the speech signals. State of the art multi-channel linear prediction methods do not take into account the spatial locations of the microphones and assume single distance compact microphone arrays. In this paper a spatially modified multi-channel autoregressive compression and coding method is proposed and successfully tested in order to adapt the standard multi-channel method to the virtual reality and immersive video conferencing applications where the microphones can be meters away from each other. The proposed method estimates the spatial distances between each microphone and the source to optimise the joint compression of the microphones covering a wide area. The results suggest that the proposed method outperforms the standard multi-channel compression and coding method when applied to the ad-hoc scenarios.

Index Terms—Ad-hoc signal processing, Immersive meetings, Multi-channel compression and coding, Virtual reality

1. INTRODUCTION

The portable digital devices such as smart phones and tablets with recording capabilities are ubiquitous and are being increasingly applied as online meeting platforms between multiple parties. Such so called ad-hoc microphone arrays [1] formed by portable recording devices are an emerging approach to enhance the quality of captured acoustics using the available resources in an acoustic enclosure [2] and are integrated into interconnected virtual reality and immersive technologies [3]. As ad-hoc microphones are randomly spread out within an unknown acoustic scene, blind spatial modification and spatial filtering of the recorded signals are important steps of ad-hoc speech and audio signal processing [4, 5].

Joint analysis of the ad-hoc microphones is challenging as the microphones are usually not synchronised, the array topology is unknown and potentially changeable and the microphones have different gains and qualities [6]. Due to such problematic limitations joint analysis of the ad-hoc microphones usually requires assumptions such as the knowledge of the number of sources [7] or the Room Impulse Responses (RIRs) at each microphone location [8].

Researchers have tried to adapt the standard array processing techniques to the ad-hoc arrays however it is shown that these adaptations require important considerations such as spatial selectivity [9] and the prior knowledge of the array geometry [10].

Multi-channel Linear Prediction (LP) is widely used for signal processing techniques and applications such as signal compression, beamforming [11] speech enhancement [12, 13] and speech recognition [14]. In the context of ad-hoc signal processing it is

important to modify the standard multi-channel LP in order to consider the wide spatial coverage of the array as well as inconsistencies in microphone gains.

It is shown that in reverberant environments applying multi-channel (7 microphones) linear prediction outperforms the single channel case [15] for modelling the reverberant speech. It is also observed that increasing the inter-channel spacing from 5cm to 35cm decreases the Itakura distance between the clean speech signal Auto Regressive (AR) coefficients and the AR coefficients obtained by applying the multi-channel LP to the microphone recordings.

This paper introduces the spatial multi-channel linear prediction and the spatial beamformer which take into account the estimated source to microphones distances without any prior information about the acoustic scene. The proposed spatial compression method is applicable to microphone arrays of arbitrary geometry with inconsistent gains. Also a spatially modified beamformer is proposed and applied for the AR modelling and compression of the source signal through the joint analysis of reverberant multi-channel ad-hoc recordings. The applied statistical distance [16] and reverberation level cues [17] proposed in this work are applicable to microphones with inconsistent gains and do not require identical dual-channel microphones [18, 19]. The target application of this research is the speech compression for the immersive meeting scenarios [20] where the microphones do not form a fully connected Wireless Acoustic Sensor Network (WASN) and the participants have their independent recording device and microphone. Under such assumptions proposed beamformers for the distributed microphones [21] are not applicable.

2. MATHEMATICAL MODEL OF AD-HOC RECORDINGS

Considering an ad-hoc scenario with M microphones of arbitrary gains [9] the recorded time frame ($x_m(n)$) of length T_f , by microphone $m \in \{1, \dots, M\}$, is mathematically modelled as

$$x_m(n) = h_m(n) * s(n) + v(n), \quad (1)$$

where $h_m(n)$ is the Room Impulse Response (RIR) for microphone m and the active source and $s(n)$ and $v(n)$ represent the source signal and the noise respectively. n is the discrete time index and m is the microphone index. It is assumed that only one speaker is active during each time frame. The matrix of the ad-hoc recordings by all the M microphones is obtained as

$$\mathbf{X}(n) = \begin{bmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{bmatrix}. \quad (2)$$

Each single channel recording $x_m(n)$, $m \in \{1, \dots, M\}$ is modeled by the AR filter as

$$x_m(n) = a_m^T x_m(n-1) + e_m(n) \quad (3)$$

where $a = [a_1, a_2, \dots, a_P]^T$ is $P \times 1$ vector of the AR coefficients and $[\cdot]^T$ denotes the matrix transpose. $x_m(n-1) = [x_m(n-1), x_m(n-2), \dots, x_m(n-P)]$ and $e_m(n)$ is the prediction error or residual [15].

In linear prediction, the AR coefficients (LP coefficients) are obtained by minimising the sum of the squared prediction error,

$$J = \sum_{n=-\infty}^{n=+\infty} e_m^2(n) \quad (4)$$

$$= \sum_{n=-\infty}^{n=+\infty} (x_m(n) - a^T x_m(n-1))^2$$

The multi-channel model of (4) for M microphones is

$$J_M = \frac{1}{M} \sum_{m=1}^M \sum_{n=-\infty}^{n=+\infty} e_m^2(n) \quad (5)$$

$$= \frac{1}{M} \sum_{m=1}^M \sum_{n=-\infty}^{n=+\infty} (x_m(n) - a_m^T x_m(n-1))^2. \quad (6)$$

It is shown that in reverberant environments the mathematical expectation of the LP coefficients ($E[a_m]$) derived from speech recordings (8) is equal to the LP coefficients derived from the source signal (b) [15].

$$s(n) = b^T s(n-1) + e_s(n) \quad (7)$$

where $s(n-1) = [s(n-1), s(n-2), \dots, s(n-P)]$

$$E[a_m] = \frac{1}{M} \sum_{m=1}^M a_m \cong b. \quad (8)$$

3. SPATIAL SPEECH COMPRESSION AND CODING

The ad-hoc microphones spatially distributed within a reverberant room are distorted by the reverberation and noise unequally and the LP coefficients derived from the reverberant recordings are also inaccurate [15]. Utilising the distance cues, facilitates a spatially modified beamformer (Figure 1) and a multi-channel LP analysis (Figure 2) that takes into account the spatial locations of the microphones and estimates the source LP coefficients more accurately.

Figure 1: Spatial DSB beamformer

Figure 2: Spatial multi-channel compression

In this paper two different approaches are proposed to obtain the source LP coefficients, 1) spatially modified multi-channel LP and 2) LP coefficients derived from a spatial beamformer output. Both methods utilise the microphone to source distance estimates and deliver one set of LP coefficients by the joint analysis of the ad-hoc recordings.

3.1. Spatial multi-channel AR modelling

In this section the standard LP analysis of the reverberant speech [15] based on the autocorrelation function [12], is spatially modified in order to estimate the clean LP coefficients more accurately.

The standard auto-correlation function $r_m(c)$ is obtained for channel m from (1) by

$$r_m(c) = E[x_m(n)x_m(n+c)], \quad c = 0, 1, 2, \dots \quad (9)$$

where E denotes the mathematical expectation. The baseline averaged autocorrelation function [12] is obtained by

$$\bar{r}(c) = \frac{1}{M} \sum_{m=1}^M r_m(c). \quad (10)$$

As it is inferred from (10) all the M autocorrelation functions are equally weighted in the averaging process (i.e. no weighting scheme is applied). The averaged autocorrelation function can be written in a more general form of a weighted average autocorrelation ($\bar{r}_w(c)$), in order to account for the source to microphone distances in the multi-channel linear error minimisation (5). Assuming that the applied weights are $\beta = \{\beta_1, \dots, \beta_M\}$ the weighted average autocorrelation function is calculated as

$$\bar{r}_w(c) = \frac{1}{\sum_{m=1}^M \beta_m} \times \sum_{m=1}^M \beta_m r_m(c), \quad (11)$$

where β_m is the weights to $r_m(c)$ and $\bar{r}_w(c)$ is the weighted average auto-correlation function. The LP filter coefficients are obtained by the Yule-Walker method (12).

$$\bar{a} = \begin{bmatrix} \bar{r}_w(0) & \dots & \bar{r}_w(P-1) \\ \vdots & \ddots & \vdots \\ \bar{r}_w(P-1) & \dots & \bar{r}_w(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \bar{r}_w(1) \\ \vdots \\ \bar{r}_w(P) \end{bmatrix}, \quad (12)$$

Assuming that the ground truth source to microphone distances for all the M microphones are $\{q_{1,s}, \dots, q_{M,s}\}$, the ideal distance weights are $\mathbf{q} = \{\frac{1}{q_{1,s}}, \dots, \frac{1}{q_{M,s}}\}$. It is observed that using $\mathbf{q}$ (where available) as the weights significantly improves the autocorrelation function

TABLE I
DISTANCE CUES SUMMARY

MSC	Only applicable to the dual-channel microphones,
Kurtosis of the LP residual	Independent of the microphone gain
Skewness of the LP residual	Independent of the microphone gain
Signal energy [22]	Sensitive to the microphone Gains and noise

estimation by (11). In other words applying the inverse of the source to microphone distances as the weight in (11) estimates the clean autocorrelation function more accurately than (10). However, the knowledge of the source to microphone distances ($\mathbf{q}$) is not usually available or retrievable and using $\mathbf{q}$ is not practical for the ad-hoc scenarios.

3.2. LP coefficients derived from a spatial beamformer output

The Delay and Sum Beamformer (DSB) is successfully applied to the ad-hoc scenarios [22] whereas other beamformers such as Minimum Variance Distortionless Response (MVDR) beamformers are shown to be impractical for dynamic ad-hoc scenarios [10, 11]. In this section the standard DSB (13) is modified based on the estimated source to microphone distances and the single channel output of the beamformer is then utilised for deriving the LP coefficients.

The baseline DSB [22] is mathematically modelled as

$$\hat{x}(n) = \frac{1}{M} \sum_{m=1}^M x_m(n - \tau_m), \quad (13)$$

where τ_m is the time delay between microphone m and the reference channel. The reference channel can be chosen randomly and obviously $\tau_m = 0$, for the reference microphone. The estimated source LP coefficients are then obtained by applying standard single channel LP to $\hat{x}(n)$ [15]. In this section the spatial DSB (Figure 2) is proposed which utilises $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_M\}$ as the weights to the microphone signals for spatial modification. The general weighted DSB is mathematically modelled as

$$\hat{x}_w(n) = \frac{1}{\sum_{m=1}^M \beta_m} \times \sum_{m=1}^M \beta_m x_m(n - \tau_m). \quad (14)$$

4. DISTANCE CUES FOR AD-HOC MICROPHONES

The Magnitude Squared Coherence (MSC) [18], the kurtosis of the LP residual signal [16] and the skewness of the LP residual signal are applicable distance cues to the ad-hoc microphones. In this research statistical cues, Kurtosis and Skewness [17], derived from the single channel LP residual signals are applied as the distance estimates. Compared with acoustic features such as Magnitude Squared Coherence (MSC) which requires dual-channel microphones of the same structure [18, 19], the applied statistical features are applicable to any microphone type.

Although it is shown that inconsistent microphone gains can be mathematically accounted for [6], the kurtosis and the skewness of the LP residual signal are independent of the microphones gains and can be applied to an array of microphones with inconsistent gains without any extra computational cost.

4.1. Kurtosis of the LP residual signal

Clean LP residual signals have distinct peaks generated by the speaker glottal cycles whereas reverberant LP residuals have the

glottal peaks followed by several peaks generated by the reverberation. This distortion affects the kurtosis of the LP residual signal [16] and it is shown that the kurtosis and distance have an nonlinear inverse relationship.

The kurtosis of $e_m(n)$ is calculated as

$$\beta_m = \frac{E\{e_m^4(n)\}}{E^2\{e_m^2(n)\}} - 3 \quad (15)$$

$$= \frac{\frac{1}{T_f} \sum_{n=1}^{T_f} (e_m(n) - \bar{e}_m(n))^4}{\left(\frac{1}{T_f} \sum_{n=1}^{T_f} (e_m(n) - \bar{e}_m(n))^2\right)^2} - 3 \quad (16)$$

4.2. Skewness of the LP residual signal

Skewness of the LP residual signal decays with the level of reverberation in the speech signal [17] or in other words microphones located closer to the speech source have higher skewness of the LP residual signals. This observation makes the skewness of the LP residual signal a suitable distance feature that can be calculated blindly for each microphone without any prior knowledge for ad-hoc scenarios.

The skewness of a random variable (e.g. LP residual signal) is defined as

$$\gamma = \frac{\mu_3}{\sigma^3} \quad (17)$$

where μ_3 is the third statistical moment and σ is the standard deviation [17].

5. RESULTS AND EVALUATION

The proposed spatial multi-channel linear prediction and the spatial beamformer are applied in order to investigate the effect of the weights and compare the results with the baseline multi-channel linear prediction [12]. Figure 3 illustrates the relationship between the proposed weights and the source to microphone distance. One microphone is moved away from a source with 0.1m step size and the proposed weights are calculated over a 200ms time frame for each setup. It is observed that the kurtosis of the LP residual signal has a larger variance compared to the skewness of the LP residual signal which leads to a more significant effect of the microphones located closer to the source in the joint analysis process.

Figure 3: The proposed weights calculated for one frame size at different distances from the source

TABLE II
EXPERIMENTAL CONFIGURATION

Sampling Frequency (f_s)	16 kHz
P	20
Reverberation time (RT_{60})	{200, 400, 600, 800} ms
Number of scenarios	250 for each RT_{60}
Room dimensions (x, y, z)	(10m, 8m, 3m)
Number of microphones (M)	4
T_f	320 (20ms)
Maximum inter-channel distance	12m

Figure 4: Itakura error vs reverberation time for the spatial multi-channel compression

Table II shows the experimental setup. Four microphones are randomly positioned in different topologies for each reverberation time whereas the source is fixed at one spot. 20 milliseconds time aligned frames are applied for the spatial multi-channel LP ((10)-(12)) with $P=20$. Room dimensions knowledge are not used in the proposed approaches.

The Itakura distance (18) is applied as the LP coefficient estimation error measurement.

$$Er_{\bar{a},b} = \left| \sum_{k=1}^P \left(\frac{\bar{a}_k}{b_k} - \log \frac{\bar{a}_k}{b_k} - 1 \right) \right| \quad (18)$$

The Itakura error ($Er_{\bar{a},b}$) [15] is calculated for the source LP coefficients and the LP coefficients obtained by the proposed and baseline methods. 250 random scenarios are simulated and the average results are presented.

Figure 4 shows that for longer reverberation times the proposed spatial multi-channel LP outperforms the baseline multi-channel LP [12] when applied to ad-hoc scenarios. Figure 5 compares the baseline DSB [22] with the weighted spatial DSB. It is concluded

Figure 5: Itakura error vs reverberation time for the spatial LP coefficients obtained by the spatial beamformer

that the weighted DSBs do not significantly improve the Itakura error.

Comparing Figure 4 and Figure 5 it is observed that the proposed spatial multi-channel LP outperforms the spatial DSB beamformer. This observation made for the ad-hoc scenarios is similar to the results from [15] which studied a linear array. The explanation of this observation can be the robustness of the LP coefficients against reverberation as the reverberation mainly affects the residual signals [16]. In other words it is recommended to average the LP coefficients rather than the speech signals as the LP coefficients are less sensitive to reverberation whereas the speech signals are highly distorted by the reflections.

The experiments of this work shows that the kurtosis and the skewness of the LP residuals are reliable source to microphones distance estimators in reverberant environment and can be applied as side information to modify the standard compression methods.

6. CONCLUSION

A spatial multi-channel speech compression and coding method and a spatially modified beamformer are proposed and successfully tested for virtual reality and immersive meeting applications. The aim of this spatial modification is to estimate the source AR coefficients more accurately in reverberant environments by considering the relative source to microphone distance estimates. The proposed methods are applicable to the ad-hoc scenarios where the microphones are arbitrarily located at unknown locations. The proposed method does not require any specific microphone structure or prior knowledge of the room dimensions. Although in this work the distance weights are only calculated once and only static acoustic scenes are simulated however by updating the weights over time, the proposed method will be applicable to dynamic scenarios with moving microphones and sources. Effect of noise on the spatial compression will be investigated in the future work.

REFERENCES

- [1] I. Himawan, I. McCowan and S. Sridharan, "Clustered Blind Beamforming From Ad-Hoc Microphone Arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661-676, May 2011.
- [2] V. Tavakoli, J. Jensen, R. Heusdens, J. Benesty and M. Christensen, "Ad hoc microphone array beamforming using the primal-dual method of multipliers," in *24th European Signal Processing Conference (EUSIPCO)*, Budapest, 2016.
- [3] E. Bastug, M. Bennis, M. Medard and M. Debbah, "Toward Interconnected Virtual Reality: Opportunities, Challenges, and Enablers," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 110-117, June 2017.
- [4] S. Gergen and R. Martin, "Estimating Source Dominated Microphone Clusters in Ad-Hoc Microphone Arrays by Fuzzy Clustering in the Feature Space," in *Speech Communication; 12. ITG Symposium*, Paderborn, Germany, 2016.
- [5] V. Tavakoli, J. Jensen, J. Benesty and M. Christensen, "A partitioned approach to signal separation with microphone ad hoc arrays," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016.
- [6] Z. Liu, Z. Zhang, L. He and P. Chou, "Energy-Based Sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, 2007.
- [7] S. Gergen, A. Nagathil and R. Martin, "Audio signal classification in reverberant environments based on fuzzy-clustered ad-hoc microphone arrays," in *International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013.
- [8] W. Wood, E. Hadad, I. Merks, B. Xu, S. Gannot and T. Zhang, "A real-world recording database for ad hoc microphone arrays," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2015.
- [9] V. Tavakoli and Benesty, "Pseudo-Coherence-based MVDR beamforming for speech enhancement with ad-hoc microphone arrays," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, South Brisbane, 2015.
- [10] N. D. Gaubitch, J. Martinez, W. B. Kleijn and R. Heusdens, "On near-field beamforming with smartphone-based ad-hoc microphone arrays," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, 2014.
- [11] Y. Zhang, D. Florencio and M. Hasegawa-Johnson, "Glottal Model Based Speech Beamforming for Ad-Hoc Microphone Arrays," in *INTERSPEECH*, Stockholm, Sweden, 2017.
- [12] Kinoshita, Delcroix and Miyoshi, "Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534-545, 2009.
- [13] M. Delcroix, T. Hikichi and M. Mioshi, "Precise dereverberation using multichannel linear prediction," *IEEE Transaction Audio, Speech, and Language processing*, vol. 15, no. 2, pp. 430-440, Feb.2007.
- [14] P. Alku and R. Saeidi, "The Linear Predictive Modeling of Speech From Higher-Lag Autocorrelation Coefficients Applied to Noise-Robust Speaker Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1606-1617, 2017.
- [15] Gaubitch, D. Ward and P. A. Naylor, "Statistical analysis of the AR modeling of reverberant speech," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, p. 4031-4039, 2006.
- [16] K. M. Hayashida, T. Nishiura, Y. Yamashita and T. Kato, "Close/distant talker discrimination based on kurtosis of linear prediction residual signals," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014.
- [17] S. Mosayyebpour, H. Sheikhzadeh, T. Gulliver and M. Esmaeili, "Single-Microphone LP Residual Skewness-Based Inverse Filtering of the Room Impulse Response,," *IEEE Transactions on Audio, Speech, and Language Processing" in IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1617-1632, 2012.
- [18] S. Vesa, " Binaural Sound Source Distance Learning in Rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1498-1507, 2009.
- [19] S. Pasha, J. Donley, C. Ritz and Y. X. Zou, "Towards real-time source counting by estimation of coherent-to-diffuse ratios from ad-hoc microphone array recordings," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, 2017.
- [20] E. Southgate and S. P. Smith, "Designing and conducting research using immersive technologies in schools: Seven observations," in *IEEE Virtual Reality Workshop on K-12 Embodied Learning through Virtual & Augmented Reality (KELVAR)*, Los Angeles, CA, 2017.
- [21] M. Connor, W. Kleijn and T. Abhayapala, "Distributed sparse MVDR beamforming using the bi-alternating direction method of multipliers," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016.
- [22] I. Himawan, I. McCowan and S. Sridharan, "Clustering of ad-hoc microphone arrays for robust blind beamforming," in *International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010.
- [23] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010.