2008

Multivariate autoregressive modelling of multichannel reverberant speech

Eva Cheng
University of Wollongong, ecc04@uow.edu.au

Ian Burnett
University of Wollongong, ianb@uow.edu.au

Christian Ritz
University of Wollongong, critz@uow.edu.au

Multivariate autoregressive modelling of multichannel reverberant speech

Abstract

Recent research in speech localization and dereverberation introduced processing of the multichannel linear prediction (LP) residual of speech recorded with multiple microphones. This paper investigates the novel use of intra- and inter-channel speech prediction by proposing the use of a multichannel LP model derived from multivariate autoregression (MVAR), where current LP approaches are based on univariate autoregression (AR). Experiments were conducted on simulated anechoic and reverberant synthetic speech vowels and real speech sentences; results show that, especially at low reverberation times, the MVAR model exhibits greater prediction gains from the residual signal, compared to residuals obtained from univariate AR models for individually or jointly modelled speech channels. In addition, the MVAR model more accurately models the speech signal when compared to univariate LP of a similar prediction order and when a smaller number of microphones are deployed.

Disciplines

Physical Sciences and Mathematics

Publication Details

E. Cheng, I. S. Burnett & C. H. Ritz, "Multivariate autoregressive modelling of multichannel reverberant speech," in International Workshop on Multimedia Signal Processing, 2008, pp. 945-949.

Multivariate Autoregressive Modelling of Multichannel Reverberant Speech

E. Cheng[#1], I. S. Burnett[*2], C. H. Ritz[#3]

[#] *School of Electrical, Computer, and Telecommunications Engineering*
University of Wollongong, Wollongong NSW Australia 2522
[1] ecc04@uow.edu.au
[3] critz@uow.edu.au

[*] *School of Electrical and Computer Engineering,*
Royal Melbourne Institute of Technology, Melbourne VIC Australia 3000
[2] ianb@uow.edu.au

***Abstract*—Recent research in speech localization and dereverberation introduced processing of the multichannel linear prediction (LP) residual of speech recorded with multiple microphones. This paper investigates the novel use of intra- and inter-channel speech prediction by proposing the use of a multichannel LP model derived from multivariate autoregression (MVAR), where current LP approaches are based on univariate autoregression (AR). Experiments were conducted on simulated anechoic and reverberant synthetic speech vowels and real speech sentences; results show that, especially at low reverberation times, the MVAR model exhibits greater prediction gains from the residual signal, compared to residuals obtained from univariate AR models for individually or jointly modelled speech channels. In addition, the MVAR model more accurately models the speech signal when compared to univariate LP of a similar prediction order and when a smaller number of microphones are deployed.**

I. Introduction

Speech recorded with multiple microphones placed in a reverberant room is subject to degradations caused by convolutive reverberation and additive background noise. Recent research that processes residual signals derived from linear prediction (LP) of the multichannel speech signals has shown performance improvements in areas such as speaker time-delay estimation and localization [1], and speech dereverberation [2][3][4]. Deploying multiple microphones in a reverberant space provides spatial diversity and signal redundancy: the speech signal is common to all recorded signals, but reverberation effects differ between channels. Thus, multichannel processing can enhance the channel/s least degraded by reverberation, or process channels together to minimise the effects of reverberation (e.g., beamforming).

Processing the recorded speech with LP can then further enhance general signal processing approaches by exploiting the speech signal characteristics. Different techniques have been proposed for linear prediction analysis of multichannel speech: Raykar et al. individually process the channels using standard linear prediction analysis [1]; Delcroix et al. [2] propose the Linear-predictive Multi-input Equalization (LIME) algorithm; Triki et al. [3] apply multichannel LP to pre-whitened speech input; and Gaubitch et al. [4] propose spatially averaged LP coefficients.

This paper proposes the use of multivariate autoregressive (MVAR) modelling, commonly used in the natural sciences, biomedicine, and economics, for multichannel speech LP: previous (univariate LP) work has not considered inter-channel prediction to exploit multi-microphone speech recordings. Experiments in this paper compare the proposed MVAR approach to current multichannel LP techniques based on a univariate autoregressive (AR) model. This paper studies the generalized multichannel LP techniques [1][4], rather than approaches derived for a particular speech application such as dereverberation [2][3].

The inter-channel prediction of MVAR takes advantage of signal redundancy between highly correlated microphone channels to derive an accurate speech signal model in reverberant environments for applications in speech enhancement, dereverberation, and localization. An additional motivation for using MVAR for speech recordings is the information potentially contained within the inter-channel prediction coefficients: derivation of the coefficients is effectively an inter-channel cross-correlation procedure, therefore studying the inter-channel prediction coefficients can yield information about the time-delay between channels (and hence source location information), in addition to information about the room reverberation characteristics.

In the remainder of this paper, Section 2 summarizes current multichannel LP approaches and presents the multivariate LP model for speech. Section 3 describes the simulated and real recordings used in the experiments, with the results presented in Section 4. Section 5 concludes this paper.

II. Linear Prediction Analysis

A. Single Channel LP

Single channel Linear Prediction (LP) is a univariate autoregressive (AR) technique, where samples in a speech channel are predicted as a weighted sum of the past P samples of that channel, and where P is the predictor order. The error (or residual) signal for each channel c ($e_c(n)$), is defined as the difference between the original ($s_c(n)$) and predicted ($\hat{s}_c(n)$) speech signals. The LP analysis procedure is given by:

$$\hat{s}_c(n) = \sum_{k=1}^{P} a_{k,c} s_c(n-k);\ e_c(n) = s_c(n) - \hat{s}_c(n) \qquad (1)$$

The prediction coefficients, $a_{k,c}$, are calculated by minimizing the square of the error signal, $e_c(n)$ over a frame of N samples. This leads to solving the set of linear equations based on the autocorrelation functions $R_c(i)$ of $s_c(n)$:

$$R_c(i) = \sum_{k=1}^{P} a_{k,c} R_c(i-k) \ \text{ where } \ R_c(i) = \sum_{n=i}^{N} s_c(n) s_c(n-i) \quad \text{for } i = 1, 2, \ldots, P. \qquad (2)$$

In this paper, Levinson-Durbin recursion [5] is used to solve for the prediction coefficients of Eq. (2), and each multichannel speech signal is then filtered with its LP model to obtain the LP residual signal, $e_c(n)$ of Eq. (1).

B. Univariate Autoregressive Multichannel LP

To extend the concepts of single channel LP to multichannel speech, Gaubitch et al. proposed an averaged (across channels) autocorrelation matrix, R_{avg}, instead of R_c in Eq. (2) [4]:

$$R_{avg}(i) = \sum_{k=1}^{P} a_{k,avg} R_{avg}(i-k) \ \text{ where } \ R_{avg} = \sum_{c=1}^{C} R_c \quad \text{for } i = 1, 2, \ldots, P. \qquad (3)$$

Levinson-Durbin recursion then solves Eq. (3) to find a LP coefficient set, $a_{k,avg}$, jointly calculated across the speech channels.

C. Multivariate Autoregressive Multichannel LP

The multichannel LP approach proposed in this paper employs multivariate autoregression (MVAR) on the multichannel speech; that is, the speech samples of a channel are predicted from P past samples of current channel and P past samples of all the other speech channels. This process is represented by:

$$e_c(n) = s_c(n) - \hat{s}_c(n) = s_c(n) - \sum_{m=1}^{C} \sum_{k=1}^{P} a_{c,m}(k) s_m(n-k) \qquad (4)$$

where each $a_{c,n}$ is a P length vector containing the intra-channel prediction coefficients (for c=m), and inter-channel linear prediction coefficients between channel c and m (for $c \neq m$); this then leads to $P \times C \times C$ MVAR prediction coefficients in total for the C channels.

Similar to univariate LP, the squared error must be minimized across all n to find the optimal matrix of prediction coefficients. However, the standard Levinson-Durbin recursion cannot be applied to multivariate (vector) prediction; rather, the Levinson-Wiggins-Robinson algorithm is one well-used MVAR extension of the single channel Levinson recursion [6]. Finally, to obtain the matrix of residual signals, a multivariate filter is required to filter each channel with the multivariate prediction coefficient matrix and all C speech channels [6].

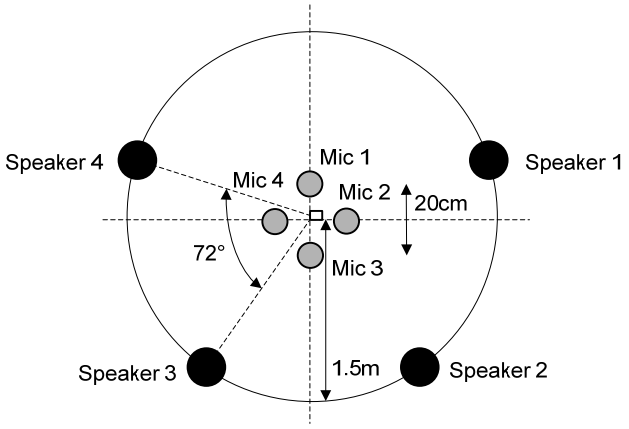

Fig. 1. Simulated recording setup

III. Reverberant Speech Recordings

To evaluate the proposed system with ideal (voiced) speech source signals for LP analysis, five English vowels ('a', 'e', 'i', 'o', 'u') of approx. 200ms in duration were synthesized using the ProSynth software, which employs a hierarchical phonological structure for speech synthesis [7]. To evaluate MVAR over a variety of speech conditions, five real speech sentences (each approx. 2s long), three female and three male, from the Australian National Database of Spoken Languages (ANDOSL) database [8] were also tested. Vowel and speech signals were sampled at 8kHz, and stored at 16 bits/sample.

To simulate spatially distributed sources, four speakers were placed in a circle of 3m in diameter and 'recorded' with four microphones placed within the circle at the centre. This experimental setup, illustrated in Fig. 1, was modelled using Allen and Berkeley's image method [9], with reverberation times (T60) ranging from anechoic (T60=0) to T60=1s. The vowels and sentences were 'played' in turn from the four source locations and 'recorded' with the four omnidirectional microphones.

IV. Results and Discussion

With the 'recorded' speech sampled at 8kHz, an LP order of P=10 was chosen for Eq. (1). To maintain near-stationary speech within an analysis frame for valid autoregressive modeling, 50% overlapped, 25ms Hamming windowed analysis frames were employed.

To evaluate the proposed system, the Itakura distance [5] and prediction gain [5] performance metrics were used. The reference LP coefficients and residual signals were obtained from the anechoic speech, to maintain aligned frame boundaries as the 'recorded' signals differ temporally from the source speech by propagation delay.

The Itakura distance is used to compare individually and jointly calculated LP coefficients. The prediction gain is used to compare the performance of the univariate and multivariate multichannel LP techniques, as it is not valid to compare the AR coefficient vectors of univariate AR with the AR coefficient matrix of MVAR.

For the synthetic vowel results presented in Section A, the metrics are averaged across the four speaker locations and five vowels, whilst the speech recordings presented in Section B average the metrics over the five sentences and four speaker locations. All graphs presented exhibit 95% confidence intervals over the mean of the performance metric, and graph legends are labeled according to the microphone number (as shown in Fig. 2) and LP technique: 'Ind' refers to channels individually modeled by univariate LP (Section IIA), 'Joint' indicates channels jointly modeled by univariate LP (Section IIB), and 'MVAR' means the MVAR technique proposed for speech in this paper (Section IIC).

A. Synthetic Vowels

1) Univariate LP Itakura Distance: Fig. 2a shows that the jointly calculated LP coefficients from synthetic vowels exhibit 0.01-0.05 lower Itakura distances than the LP coefficients derived from the individually modelled speech channels, with the difference increasing with greater T60. With the low range of distance values exhibited in Fig. 2a, these differences represent up to approx. 10% of the metric value. The results in Fig. 2a confirm the findings of [4]: compared to individually modelled channels, LP coefficients jointly calculated from a synthetic vowel better match the set of coefficients obtained from clean speech.

2) Prediction Gain: In Fig. 2b, compared to the individually modelled LP, jointly calculating the LP coefficients from synthetic vowels exhibits little increase in prediction gain for all T60, despite the lower Itakura distances shown in Fig. 2a. In contrast, Fig. 2c shows MVAR to be more robust to reverberation with a consistently higher prediction gain across all T60 (especially for T60 less than 200ms), compared to the univariate AR. MVAR exhibits at least 16dB increase in gain for T60=0.1, then rapidly decreasing to about 5dB increase at higher T60. The consistently higher prediction gain exhibited by MVAR in Fig. 2c shows that the MVAR technique better predicts the speech signal in reverberant conditions: less energy in the residual signal signifies less prediction error. Lastly, the similar shapes of the curves between univariate AR and MVAR in Fig. 2c suggest that univariate AR and MVAR respond similarly to increasing reverberation.

3) Increased Univariate Prediction Order: The MVAR multichannel LP model has an increased prediction order compared to univariate LP, due to the inter-channel spatial prediction; to ensure that the improved performance of the MVAR in Fig. 2c is not due to the higher prediction order, an increased univariate AR (temporal) prediction order of $P \times C$ was tested. As shown in Fig. 3, this increased univariate LP order showed an approx. 3dB increase in jointly modelled prediction gain across all T60, compared to results shown in Fig. 2c. Although, this improved performance still lagged the MVAR results by at least 2dB for longer T60, with the MVAR model still showing up to approx. 13dB gain improvement for T60 less than 200ms.

4) Reduced Number of Microphones: To explore the effect of reducing the number of microphones used, Fig. 4 depicts

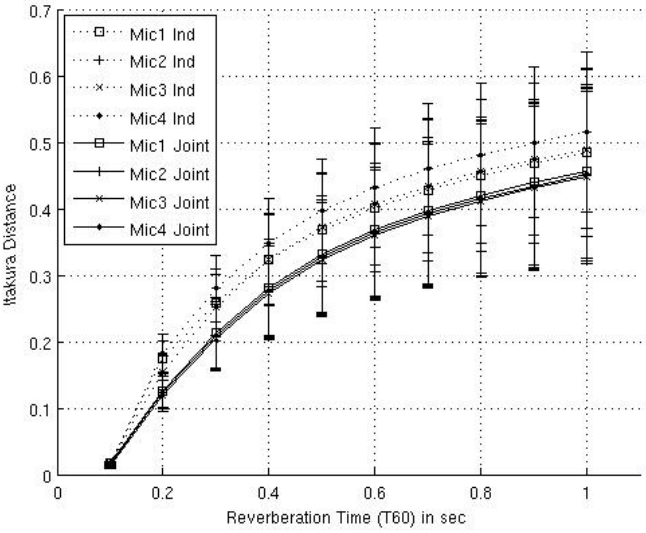

Fig. 2a. Itakura distance: univariate AR vs. Joint AR

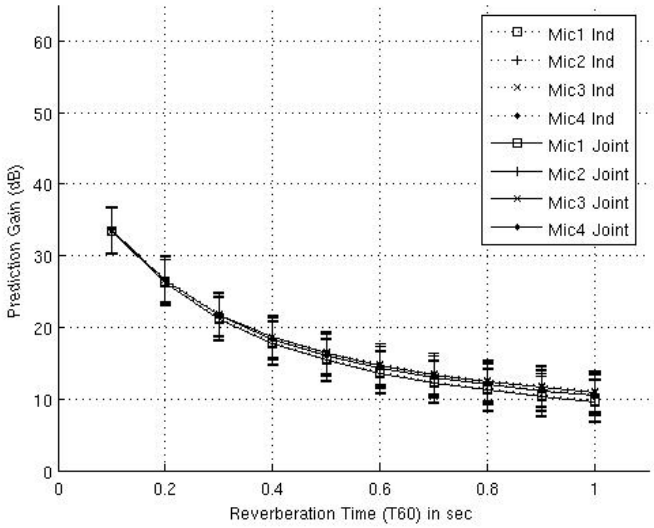

Fig. 2b. Prediction gain: univariate AR vs. Joint AR

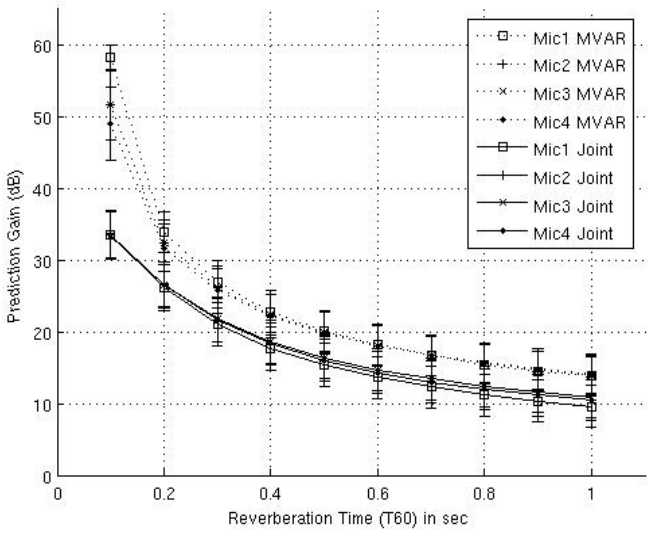

Fig. 2c. Prediction gain: Joint AR vs. MVAR

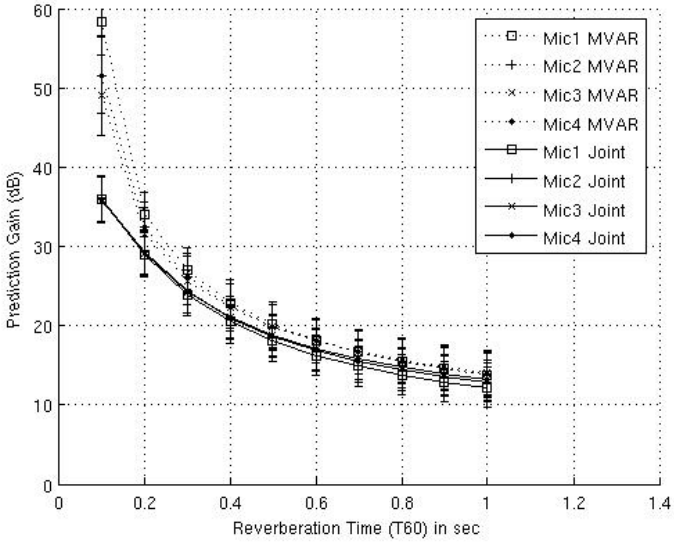

Fig. 3. Prediction gain: Joint AR (P=40) vs. MVAR (P=10)

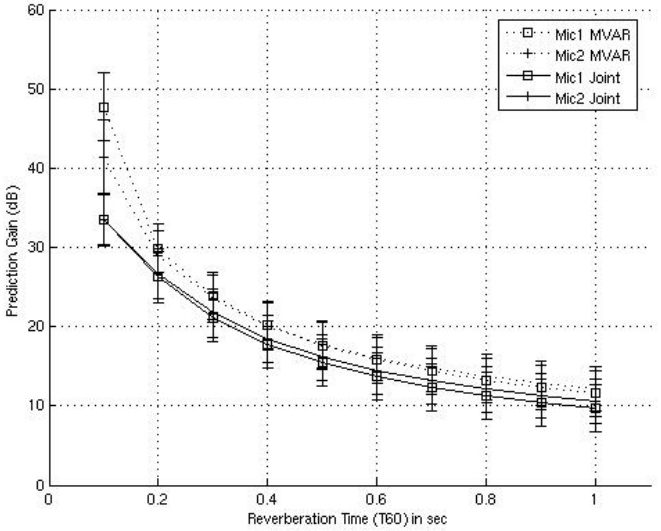

Fig. 4. 2-Channel Prediction gain: Joint AR vs. MVAR

the results from jointly modelled univariate LP and MVAR using two microphones only (Mic 1 and Mic2 in Fig. 1). Compared to Fig. 2c, Fig. 4 clearly shows that the performance of MVAR is degraded at lower reverberation times (less than 200ms), with an approx. 11dB drop in prediction gain at T60=0.1s, 5dB decrease at T60=0.2s, and up to 2dB drop at higher T60. The decrease in performance for jointly modelled univariate AR is much less marked, with 1-2dB drop in prediction gain across all T60, compared to Fig. 2c. Nonetheless, MVAR still outperforms jointly modelled univariate AR at all reverberation times, with approx. 2-15dB greater prediction gain in Fig. 4, especially at T60 less than 200ms.

B. Real Speech Sentences

1) Univariate LP: Figs. 5a and 5b show similar trends between the results obtained from synthetic vowels and real speech signals. Compared to the individually modelled channels, the jointly calculated univariate AR coefficients in Fig. 5a exhibit between 0.01-0.03 lower Itakura distances (approx. 10% of the metric value), and there is little statistically significant difference in prediction gain from individually or jointly modelled univariate LP in Fig. 5b.

2) Prediction Gain: Similar to the trends seen in Fig. 2c for the synthetic vowels, compared to the prediction gain from univariate LP, Fig. 5c illustrates consistent robustness against increasingly reverberant speech using MVAR. MVAR exhibits at least 14dB (at T60=0.1) and approx. 5dB (at higher T60) increase in prediction gain. However, for the univariate AR and MVAR LP approaches, the prediction gain becomes negative for T60 larger than 600ms and 800ms, respectively; this suggests that the LP technique is not well suited for real speech in highly reverberant conditions. But, the MVAR technique again shows increased robustness against the effects of highly reverberant degradation, with the prediction gain in Fig. 5c becoming negative at T60 200ms longer than the (individually and jointly modelled) univariate AR.

V. Discussion and Conclusion

This paper proposed the use of a multivariate autoregressive (MVAR) multichannel linear prediction (LP) model for reverberant speech. The proposed approach is compared to current multichannel speech linear prediction techniques that employ the standard univariate autoregressive (AR) LP approach, which either individually model each speech channel or derive a jointly calculated set of prediction coefficients from individually modeled channels.

The experiments in this paper were conducted on simulated recordings of synthetic speech vowels and real speech sentences in a room modeled across a range of reverberation times. Results for univariate LP showed that, in comparison to individually modeled channels of speech, LP coefficients jointly calculated across the channels more accurately match the 'ideal' set of coefficients (as obtained from anechoic signals) for both real speech sentences and synthetic vowels. However, the prediction gains are comparable between the individually and jointly modeled univariate AR models. In contrast, compared with univariate AR approaches, the proposed MVAR model exhibited significant increases in prediction gain of approx. 5-16dB (synthetic vowels) and 5-14dB (real speech sentences) across the tested reverberation times. Thus, compared to the univariate AR, MVAR is not only more robust to reverberation but also to the voiced/unvoiced and low energy signal segments inherent in real speech.

Thus, the results presented in this paper suggest that MVAR, which performs intra-channel *and* inter-channel LP, takes greater advantage of signal redundancy and spatial diversity from multi-microphone reverberant speech, compared to the univariate LP techniques. The authors are currently investigating the information contained within the MVAR inter-channel prediction coefficients. In particular, since MVAR improves the estimation accuracy of speech models in reverberant conditions, it is expected that MVAR can lead to an improvement in applications of LP to multi-microphone speech processing.

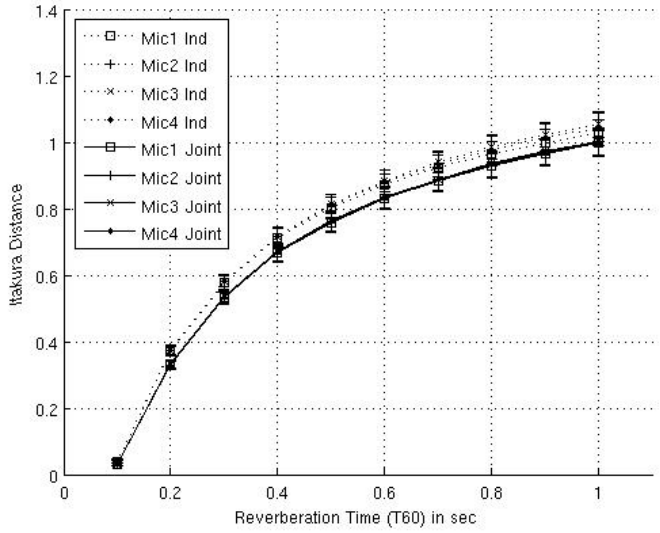

Fig. 5a. Itakura distance: univariate AR vs. Joint AR

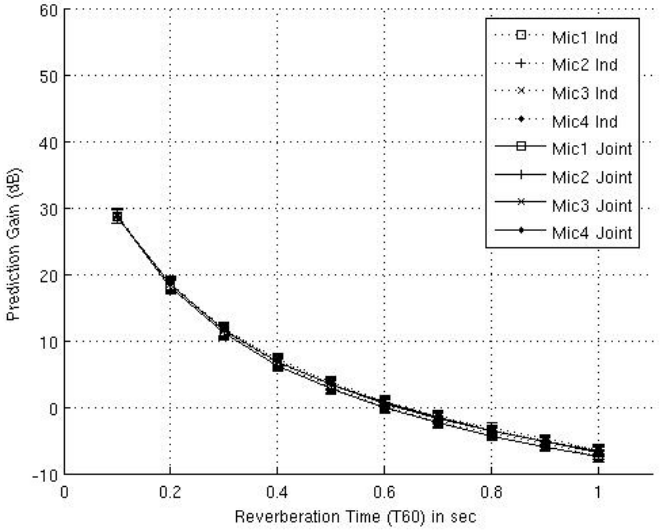

Fig. 5b. Prediction gain: univariate AR vs. Joint AR

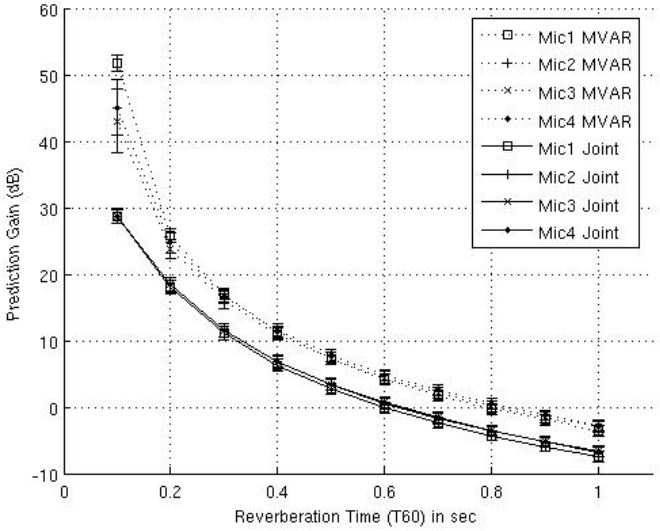

Fig. 5c. Prediction gain: Joint AR vs. MVAR

REFERENCES

[1] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Trans. Speech and Audio Proc.*, vol. 13, no. 5, pp. 751-761, Sept. 2005.

[2] M. Delcroix, T. Hikichi, M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, and Lang. Proc.*, Vol.15, No.2, pp.430-440, Feb.2007.

[3] M. Triki, D. T. M. Slock, "AR source modelling based on spatiotemporally diverse multichannel outputs and application to multimicrophone dereverberation," in *Proc. 15th Int. Conf. on DSP*, pp. 195-198, July 2007.

[4] N. Gaubitch, D. B. Ward, P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *JASA*, Vol. 120, No. 6, pp. 4031-4039, Dec. 2006.

[5] J. R. Deller Jr., J. G. Proakis, J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, New York: Macmillan, 1993.

[6] R. A. Wiggins, E. A. Robinson, "Recursive solution to the multichannel filtering problem," *J. Geophys. Res.*, vol. 70, pp. 1885-1891, 1965.

[7] (2008) ProSynth: All Prosodic Speech Synthesis. [Online]. Available: http://www-users.york.ac.uk/~lang19/

[8] (2008) Australian National Database of Spoken Language. [Online]. Available: http://andosl.anu.edu.au/andosl/

[9] J. A. Allen, D. A. Berkeley, "Image Method for Efficiently Simulating Small-Room Acoustics," *JASA*, vol. 65, no. 4, pp. 943-950, April 1979.