2007

Time delay estimation of reverberant meeting speech: on the use of multichannel linear prediction

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

I. Burnett
Faculty of Informatics, University of Wollongong, ianb@uow.edu.au

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

Time delay estimation of reverberant meeting speech: on the use of multichannel linear prediction

Abstract

Effective and efficient access to multiparty meeting recordings requires techniques for meeting analysis and indexing. Since meeting participants are generally stationary, speaker location information may be used to identify meeting events e.g., detect speaker changes. Time-delay estimation (TDE) utilizing cross-correlation of multichannel speech recordings is a common approach for deriving speech source location information. Research improved TDE by calculating TDE from linear prediction (LP) residual signals obtained from LP analysis on each individual speech channel. This paper investigates the use of LP residuals for speech TDE, where the residuals are obtained from jointly modeling the multiple speech channels. Experiments conducted with a simulated reverberant room and real room recordings show that jointly modeled LP better predicts the LP coefficients, compared to LP applied to individual channels. Both the individually and jointly modeled LP exhibit similar TDE performance, and outperform TDE on the speech alone, especially with the real recordings.

Disciplines

Physical Sciences and Mathematics

Publication Details

E. Cheng, I. S. Burnett & C. H. Ritz, "Time delay estimation of reverberant meeting speech: on the use of multichannel linear prediction", in International Conference on Signal Image Technology & Internet Based Systems (SITIS '07), 2007, pp. 494-500.

Third International IEEE Conference on Signal-Image Technologies and Internet-Based System

Time Delay Estimation of Reverberant Meeting Speech: On the Use of Multichannel Linear Prediction

E. Cheng, I. S. Burnett, C. Ritz
Whisper Laboratories
School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong NSW Australia 2522
[ecc04, ianb, critz]@uow.edu.au

Abstract

Effective and efficient access to multiparty meeting recordings requires techniques for meeting analysis and indexing. Since meeting participants are generally stationary, speaker location information may be used to identify meeting events e.g., detect speaker changes. Time-delay estimation (TDE) utilizing cross-correlation of multichannel speech recordings is a common approach for deriving speech source location information. Recent research improved TDE by calculating TDE from linear prediction (LP) residual signals obtained from LP analysis on each individual speech channel. This paper investigates the use of LP residuals for speech TDE, where the residuals are obtained from jointly modeling the multiple speech channels. Experiments conducted with a simulated reverberant room and real room recordings show that jointly modeled LP better predicts the LP coefficients, compared to LP applied to individual channels. Both the individually and jointly modeled LP exhibit similar TDE performance, and outperform TDE on the speech alone, especially with the real recordings.

1. Introduction

Multiparty meetings occur in many government, business, research, and educational environments. Recent research has focused on techniques for efficient and effective access to offline meeting recordings [1]. Analysis of meeting events is fundamental to offline access of the recordings, and Lathoud et al. proposed the use of speaker location information for meeting speech segmentation [2]. Meeting participants generally remain stationary and thus speaker location information can be used to analyze the meeting events for subsequent indexing and segmentation.

Speech, the dominant audio source in a meeting, may be localized using a number of techniques. Time-Delay Estimation (TDE) is a popular technique for deriving speech source location information: robustness to room acoustic effects common to meeting environments, such as reverberation and background noise, may be mitigated through frequency-domain weighting [3]. The application of weighted TDE defines the Generalized Cross Correlation (GCC) [3]. One particular form of weighting, GCC with Phase Transform (GCC-PHAT), has been shown to reliably derive TDE from reverberant speech. Recent research has achieved more accurate TDE through applying GCC to the speech linear prediction (LP) residual, compared to GCC-PHAT on the original multichannel speech [4]. These approaches, however, do not jointly model the LP between channels, as recently used for multichannel dereverberation of speech [5]. This paper proposes to combine these two areas of research to investigate the use of joint LP models for TDE. The proposed approach is compared to individually optimized (on a per-channel basis) LP and using the multichannel speech alone for TDE.

In the remainder of this paper, Section 2 outlines the proposed system of using a multichannel LP model front-end to GCC-based TDE. Section 3 describes the simulated and real meeting recordings used in experiments. The results are presented and analyzed in Section 4, with Section 5 concluding this paper.

2. Proposed System

Fig. 1 illustrates the proposed paradigm of using multi-channel linear prediction (LP) analysis on meeting speech (recorded with a microphone array) as a front-end to time-delay estimation utilizing GCC techniques. In the proposed system, the meeting

978-0-7695-3122-9/08 $25.00 © 2008 IEEE
DOI 10.1109/SITIS.2007.96

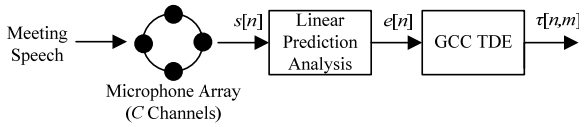


Fig. 1. Proposed approach

consists of five participants equally spaced in a circle of 3m in diameter. The meeting speech is then recorded by four microphones placed in the centre of the circle, as illustrated in Fig. 2.

2.1. Single Channel Linear Prediction

Since speech is the dominant audio source in multiparty meetings, Linear Prediction (LP) is employed for speech analysis in the proposed system. In LP, samples in the speech signal are predicted as a weighted sum of the past P samples, 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 signal. The LP analysis procedure is mathematically represented as:

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

The summing weights, $a_{k,c}$, known as linear prediction coefficients, are calculated to minimize the error signal, $e_c[n]$, energy, $E_c[n]$:

$$E_c[n] = \sum_{n=-\infty}^{\infty} {e_c}^2[n] = \sum_{n=-\infty}^{\infty} \left[s_c[n] - \sum_{k=1}^{P} a_{k,c} s_c[n-k] \right]^2 \quad (2)$$

Eq. (2) is minimized by setting $\partial E_c / \partial a_{k,c}$ for $k = 0, 1, 2, \ldots, P$, which reduces to the linear equation set:

$$\sum_{n=-\infty}^{\infty} s_c[n-i] s_c[n] = \sum_{k=1}^{P} a_{k,c} \sum_{n=-\infty}^{\infty} s_c[n-i] s_c[n-k] \quad (3)$$

for $i = 0, 1, 2, \ldots, P$.

Using the autocorrelation function $R_c[i]$ of $s_c[n]$, Eq. (3) can be reduced to (where N is length of the analysis window):

$$R_c[i] = \sum_{k=1}^{P} a_{k,c} R_c[i-k], \text{ where}$$

$$R_c[i] = \sum_{n=i}^{N} s_c[n] s_c[n-i] \text{ for } i = 1, 2, \ldots, P. \quad (4)$$

2.2. Multichannel Linear Prediction

To extend the concepts of single channel LP to multiple speech channels, Gaubitch et al. proposed the use of an averaged (across channels) autocorrelation matrix, R_{avg}, instead of R_c in Eq. (4) [6]:

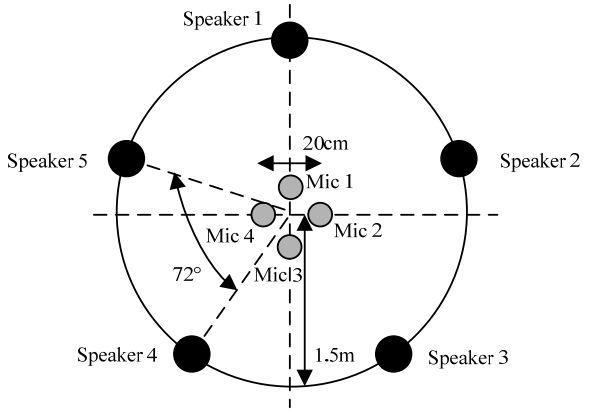


Fig. 2. Meeting room setup

$$R_{avg}[i] = \sum_{k=1}^{P} a_{k,avg} R_{avg}[i-k]$$

$$\text{where } R_{avg} = \sum_{c=1}^{C} R_c \text{ for } i = 1, 2, \ldots, P. \quad (5)$$

This paper adopts the approach in [6] to implement multichannel LP for the purposes of TDE. The Levinson-Durbin recursion algorithm is used to find the solutions of Eqs. (4) and (5) to find $a_{k,c}$ and a_k for the individual and joint LP models, respectively. Each of the multichannel speech signals is then filtered with the (individual or joint) LP model to obtain the LP residual signal, following Eq. (1).

An alternative technique to jointly model LP across multiple channels is to average the Line Spectral Frequencies (LSFs), where LSFs are an alternative representation of $a_{k,c}$. Eq. (1) can be expressed in the z-domain as:

$$A_c[z] = 1 + \sum_{k=1}^{P} a_{k,c} z^{-k} = \frac{[P_c[z] + Q_c[z]]}{2} \quad (6)$$

where $P_c[z]$ and $Q_c[z]$ are the sum and difference equations:

$$P_c[z] = A_c[z] + z^{-(P+1)} A_c[z^{-1}]$$
$$Q_c[z] = A_c[z] - z^{-(P+1)} A_c[z^{-1}] \quad (7)$$

The LSFs are defined as the polynomial roots of $P_c[z]$ and $Q_c[z]$ in Eq. (7). The LSF representation of the LP coefficients is widely used in speech coding e.g., for interpolating the LP coefficients, due to the robustness of the LSFs to quantization noise, where other representations of the LP coefficients can result in filter instability.

It is for these reasons that this paper proposes averaging the LSFs obtained from each channel as an alternative method to form the jointly modeled LP

coefficients. The roots of $P_c[z]$ and $Q_c[z]$ are found by Chebyshev polynomials methods [7], and averaged across the channels to form the averaged LSFs. The averaged LSFs are then converted back to a_k, using Eq. (6), for subsequent filtering to obtain the LP residuals using Eq. (1).

Finally, the computational complexities of averaging autocorrelation matrices and LSFs are comparable, to enable fair comparisons between the two methods.

2.3. Time-Delay Estimation

Generalized Cross Correlation (GCC) is a technique commonly applied to deriving TDE from two microphone channels [3]. Mathematically, GCC is given by:

$$\hat{G}_{X_1X_2}[k] = \frac{X_1[k] \cdot X_2^*[k]}{W[k]} \tag{8}$$

where the Discrete Fourier Transforms (DFT) of multichannel signals $x[n]$ are denoted by $X[k]$, and the frequency-domain weighting function, $W[k]$, is chosen depending on the signal and noise characteristics.

Using the Inverse Discrete Fourier Transform (IDFT), the phase correlation function is given by:

$$\hat{R}_{12}[\tau] = \text{IDFT}\left(\hat{G}_{X_1X_2}\right) \tag{9}$$

The TDE, $\hat{\tau}_{12}$, is calculated as the maximum of :

$$\hat{\tau}_{12} = \arg\max_{\tau} \hat{R}_{12}[\tau] \tag{10}$$

To minimize erroneous TDE values, the search range of delays is constrained to an interval $-D \le \hat{\tau}_{12} \le D$, where D is generally determined by the physical arrangement of the microphones.

The frequency-domain weighting function, $W[k]$, shown to be most robust to reverberant speech with low levels of noise is the PHAse Transform (PHAT), which leads to the GCC-PHAT technique [3]:

$$W[k] = \left| X_1[k] \cdot X_2^*[k] \right| \tag{11}$$

In this paper, GCC-PHAT is applied to the reverberant speech, while simple cross-correlation (CC), or GCC with $W[k] = 1$ for all frequencies, is applied to the LP residual to extract the TDE. The GCC-PHAT does not offer an advantage to LP residual signals since the PHAT weighting flattens the cross-spectrum, and the spectrum of LP residual signals is relatively flat by nature.

To apply TDE to multiple channels, GCC is calculated for each channel pair. In this paper, four microphones are deployed (see Section 3), which defines six possible microphone pairs and thus six TDE calculations.

3. Meeting Recordings

Five loudspeakers equally spaced in a circle of 3m in diameter simulated active meeting participants. Illustrated in Fig. 2, the recording setup was modeled using Allen and Berkeley's image method [7], with reverberation times (T60) from anechoic (T60 = 0) to T60 = 1 second. Most office spaces generally exhibit a reverberation time of 300 ms.

To evaluate the proposed system with ideal (voiced) speech source signals for LP analysis, the five English vowels ('a', 'e', 'i', 'o', 'u') of approx. 200ms in duration were synthesized using the ProSynth sofware, which employs a hierarchical phonological structure for speech synthesis [6]. Vowels were sampled at 16kHz, and stored at 16 bits/sample.

To simulate a meeting using the image method room model, the vowels were 'played' from the five source locations and 'recorded' with the four omnidirectional microphones, as defined in the room model of Fig. 2. Recordings were then made in a real reverberant acoustic environment of approx. 300ms reverberation time with background noise. The synthetic vowels were played in turn from the five loudspeakers (Genelec 1029A) and recorded by omnidirectional microphones (RØDE NT2A) arranged to match the room model and Fig. 2.

4. Results

To ensure real-time updates to the TDE are viable with the system proposed in this paper, 32ms Hamming windowed analysis frames are employed with 50% overlap between adjacent frames. As detailed in Section 3, the recorded speech is sampled at 16kHz, which leads to an LP order of $P = 21$ for Eq. (1).

To evaluate the proposed system, a number of performance metrics are used. All graphs presented in this section exhibit 95% confidence intervals over the specified mean of the following performance metrics:

• *Itakura distance* shows the deviation between LP autocorrelation coefficients under test, $\hat{a}_k$, and the clean speech coefficients a_k (obtained from the anechoic speech in this paper):

$$d_I = \log_{10}\left(\frac{\hat{a}_k \mathbf{R} \hat{a}_k}{a_k \mathbf{R} a_k}\right) \tag{12}$$

where $\mathbf{R}$ is the autocorrelation matrix. Thus, the smaller the Itakura distance, the closer the estimated LP autocorrelation coefficients are to the ideal case.

• *Prediction gain* is the ratio of the anechoic signal energy to the LP residual energy. Thus, the larger the prediction gain, the more accurately the LP models the vocal tract, since the residual energy is low.

• *TDE Root Mean Squared Error* (RMSE) indicates the mean square error of the TDE under study from the ground truth time delay, which is known from the microphone and speaker configuration (see Fig. 2):

$$\text{TDE RMSE} = \sqrt{\frac{1}{M \times N} \sum_{m=1}^{M} \sum_{x=1}^{X} \left(\hat{\tau}[x,m] - \tau[x,m] \right)^2} \quad (13)$$

where M is the number of possible microphone pairs, $\hat{\tau}[x,m]$ is the TDE, and $\tau[x,m]$ is the ground-truth time-delay. Since it is known that the synthetic vowels are voiced and minimally time-varying, the TDE RMSE metric is averaged across time, where X in Eq. (13) is the number of frames in the signal. Thus, the lower the RMSE, the more accurate and reliable the time delay estimation. For the room modeling results below (Sections 4.1 and 4.2), the results are averaged across the five synthetic vowels to evaluate the TDE performance across increasing reverberation time and also to evaluate the system performance with different voiced signals.

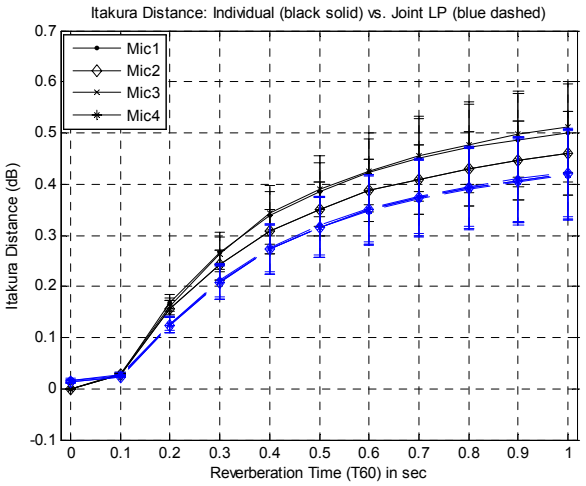
In the following sections, the Itakura distance and prediction gain performance metrics are utilized to compare the performances of TDE calculated from individually and jointly modeled LP residuals.

4.1. Autocorrelation Matrix Averaging

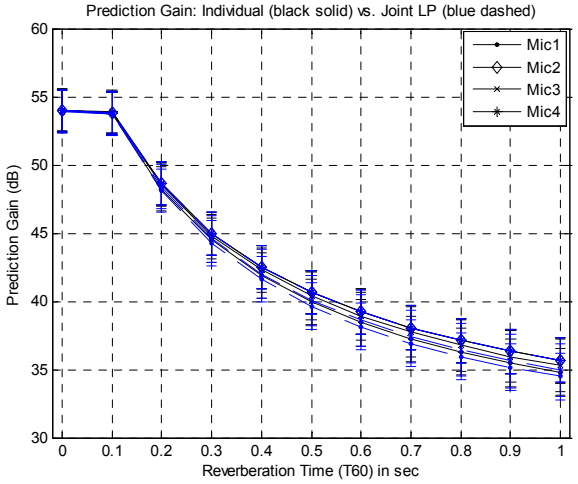
Fig. 3a shows the Itakura distance for the individually modeled microphone channels (solid lines), and for the joint LP model (dashed lines). It is clear that the jointly modeled LP model consistently outperforms the individual models with a lower Itakura distance across all reverberation times. These results confirm the statistical analyses and simulations of [6]: for a synthetic vowel signal, the joint LP model derives LP autocorrelation coefficients, a_k, that better match the ideal set of coefficients.

In contrast, Fig. 3b illustrates the prediction gain for the four microphone channels, individually (solid line) and jointly (dotted line) modeled. Although the jointly modeled LP coefficients better match the ideal set of coefficients (see Fig. 3a), when filtered with each channel of the reverberant speech to obtain the LP residual, there is little difference shown by either LP model in the prediction gain.

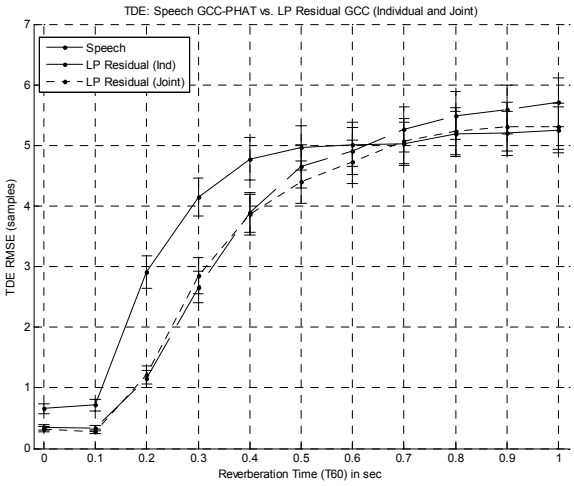
Fig. 3c illustrates the TDE performance from the reverberant speech GCC-PHAT and the individually modeled LP residual GCC. It can be clearly seen that



(a) Itakura distance: individual vs. joint LP

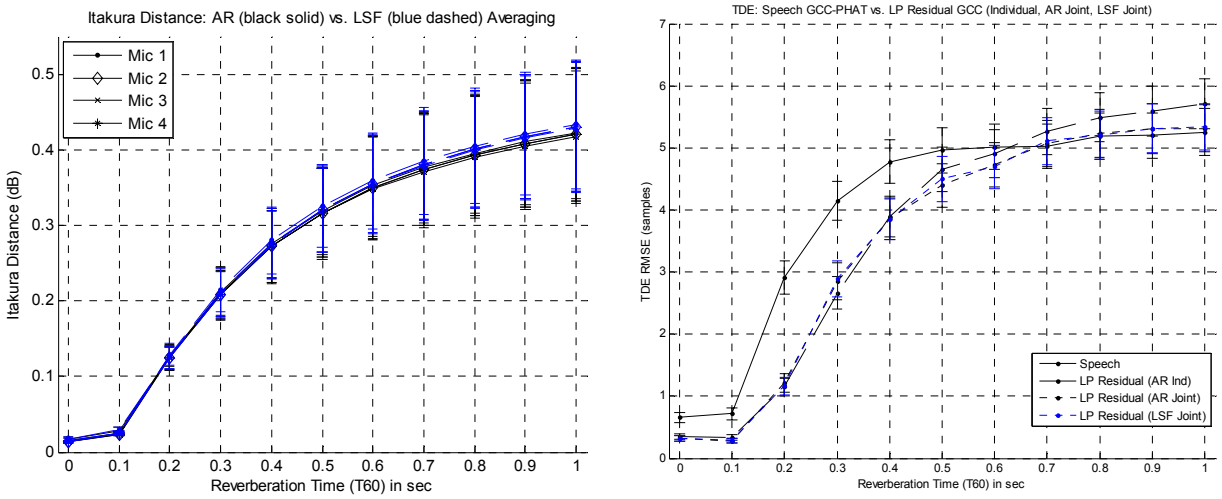


(b) Prediction gain: individual (solid line) vs. joint (dotted line) LP residual



(c) TDE RMSE: individual vs. joint LP residual

Fig. 3. Synthetic vowel simulation results



(a) Itakura distance: AR vs. LSF (b) TDE RMSE: Speech vs. Joint LP Residual (AR and LSF)

Fig. 4. Joint LP modeling: AR vs. LSF averaging

for reverberation times less than 600ms, the LP residual provides a more reliable TDE vector (across the six channel pairs, averaged in Fig. 3c) with a consistently lower TDE RMSE. As reverberation increases, however, the speech GCC-PHAT TDE exhibits slightly lower RMSE over the LP residual GCC (both individually and jointly modeled). At higher reverberation times, although the jointly modeled LP coefficients are extracted accurately compared to the individually modeled channels (see Fig. 3a), upon filtering the LP coefficients with each reverberant channel the residual can contain significant amounts of reverberation [6]. As is the case with speech, reverberation can introduce erroneous peaks into the GCC function which in turn lead to erroneous TDE.

Fig. 3c also compares the TDE RMSE from the individually (dashed line) and jointly (dotted line) modeled LP residuals. It can be seen that the jointly modeled LP increasingly improves the TDE reliability over the individually modeled channels as reverberation time increases past 400ms. However, the improvement is less then one sample in resolution.

The results in Fig. 3 suggest that in a simulated reverberant environment, while more accurately modeling the speech LP coefficients, the increased computational complexity for the jointly modeled LP model does not lead to a significant improvement in the TDE accuracy.

4.2. Line Spectral Frequencies Averaging

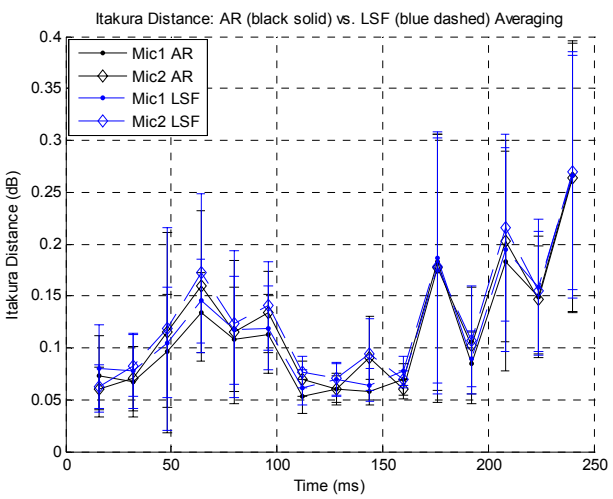
Fig. 4a shows the comparison between the Itakura distances of the joint LP models obtained by averaging autocorrelation matrices (solid line) and averaging the LSFs (dotted line). Across the simulated reverberation times, it can be seen that the LSF averaging performance is comparable to that of autocorrelation matrix averaging.

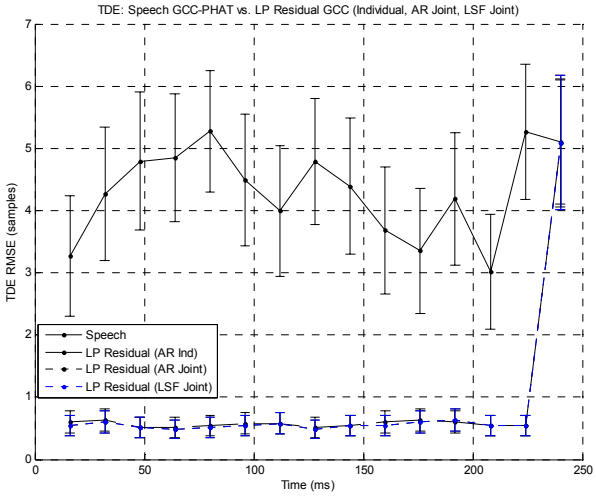
Fig. 4b depicts the TDE RMSE for the speech GCC-PHAT, and GCC of the LP residual obtained by both jointly modeled techniques. The comparable performances of the two averaging techniques shown in Fig. 4a are reciprocated with TDE reliability. The TDE performance of the two jointly modeled LP techniques is comparable to, or better than individually modeled LP and speech GCC-PHAT for reverberation times less than or greater then 400ms, respectively. Similar to the results in Fig. 3c, speech GCC-PHAT performs best at reverberation times greater then 700ms.

With similar TDE results exhibited by both the autocorrelation and LSF averaging, the results in Fig. 4 suggest that joint modeling, for both the tested methods, only result in more reliable TDE over TDE from individually modeled LP speech residuals at higher reverberation times.

4.3. Real Reverberant Recordings

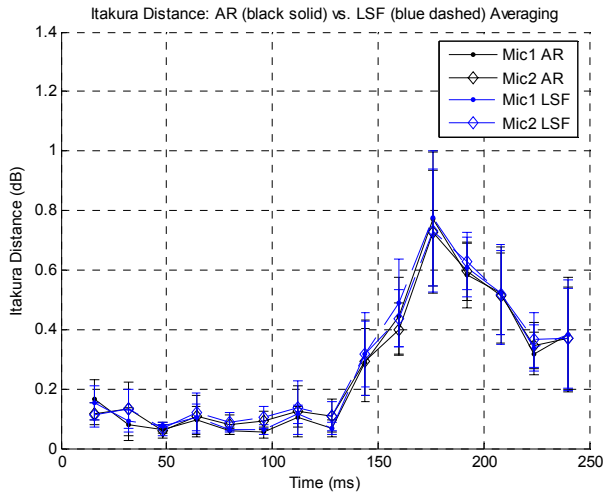
Fig. 5 shows the results from recording the ‘e’ synthetic vowel averaged over the five speaker positions, plotted across time. Similarly, Fig. 6 shows the results from recording the ‘o’ synthetic vowel. Although only the results from these two of the five synthetic vowels and two of the four microphones are presented here for brevity, the other three vowels and microphones exhibited similar trends. Both Figs. 5 and

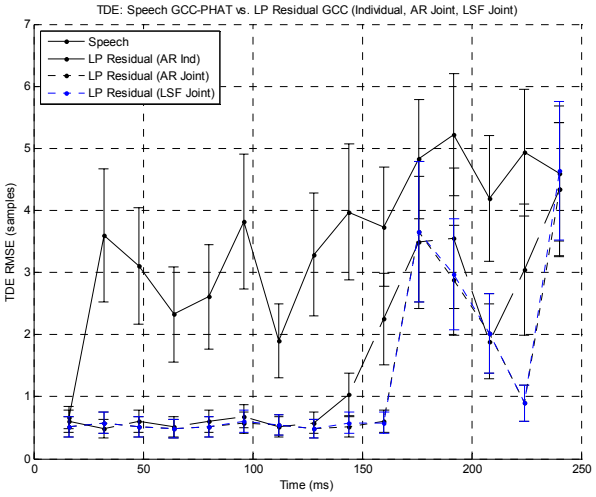




(a) Itakura distance: AR vs. LSF (b) TDE RMSE: speech vs. joint LP residual (AR and LSF)

Fig. 5. Joint LP modeling for real recording of 'e': AR vs. LSF averaging





(a) Itakura distance: AR vs. LSF (b) TDE RMSE: speech vs. joint LP residual (AR and LSF)

Fig. 6. Joint LP modeling for real recording of 'o': AR vs. LSF averaging

6 show that the performances of the autocorrelation and LSF averaging techniques are almost identical.

Figs. 5b and 6b, however, show a marked performance improvement for TDE accuracy from the LP residual (individually or jointly modeled), compared to GCC-PHAT on the speech alone. These results with real recordings confirm the findings of [4]. The jointly modeled LP residual (either AR or LSF averaged) does not significantly outperform the individually modeled LP residual, although a slight performance improvement can be seen with the 'o' vowel in Fig. 6b. The improved performance of the LP residual TDE (individually and jointly modeled) compared to speech GCC-PHAT is much more significant in a real acoustic environment compared to the theoretical simulations: this can be seen by comparing the results of Fig. 4b to those in Figs. 5b and 6b. The results in Figs. 5b and 6b clearly show that the LP residual TDE is more robust to a real reverberant acoustic environment with background noise, than the speech GCC-PHAT.

5. Conclusion

This paper studied the use of multichannel linear prediction for time-delay estimation (TDE) of reverberant speech. Two techniques for multichannel linear prediction were implemented: averaging the autocorrelation matrices, and line spectral frequencies (LSFs) across the speech channels.

The simulations in this paper were conducted on synthetic vowels in a modeled room and real recordings in a reverberant room with background noise. Results showed that jointly modeled LP coefficients better match the ideal set of LP coefficients compared to individually modeling the multiple speech channels alone. However, there is little performance gain between TDE from individually or jointly modeled LP residuals; the reasons for this are currently being investigated with both simulated and real reverberant environments. Furthermore, the two joint LP modeling techniques studied in this paper, namely, the averaged autocorrelation matrices and LSFs, perform comparably in both the simulated and real reverberant room. Nonetheless, TDE calculated from the LP residual from either technique significantly outperform the speech TDE in the real recordings. This suggests that extracting TDE from the LP residual (either individually or jointly modeled) is the most robust technique for TDE in real reverberant environments.

6. References

[1] S. Tucker and S. Whittaker, "Accessing Multimodal Meeting Data: Systems, Problems and Possibilities," in *LNCS*, vol. 3361, pp. 1-11, Springer-Verlag, Berlin, 2005.

[2] G. Lathoud, I. McCowan, "Location Based Speaker Segmentation," in proc. *ICASSP*, Hong Kong, pp. 176-179, April 2003.

[3] C. Knapp, G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoust., Speech, and Signal Proc.*, Vol. ASSP-24, No. 4, pp. 320-327, Aug. 1976.

[4] V. C. Raykar, et al., "Speaker Localization Using Excitation Source Information in Speech," IEEE Trans. Speech and Audio Proc., vol. 13, no. 5, pp. 751-761, Sept. 2005.

[5] M. Delcroix, T. Hikichi, M. Miyoshi, "Precise Dereverberation using Multichannel Linear Prediction," *IEEE Trans. Audio, Speech, and Language Proc.*, Vol. 15, No. 2, pp. 430-440, Feb. 2007.

[6] 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.

[7] P. Kabal, R. P. Ramachandran, "The Computation of Line Spectral Frequencies using Chebyshev Polynomials," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 34, No. 6, pp. 1419-1426, Dec. 1986.

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

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