2006

# Using spatial audio cues from speech excitation for meeting speech segmentation

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

# Using spatial audio cues from speech excitation for meeting speech segmentation

## Abstract

Multiparty meetings generally involve stationary participants. Participant location information can thus be used to segment the recorded meeting speech into each speaker's 'turn' for meeting 'browsing'. To represent speaker location information from speech, previous research showed that the most reliable time delay estimates are extracted from the Hubert envelope of the linear prediction residual signal. The authors' past work has proposed the use of spatial audio cues to represent speaker location information. This paper proposes extracting spatial audio cues from the Hubert envelope of the speech residual for indicating changing speaker location for meeting speech segmentation. Experiments conducted on recordings of a real acoustic environment show that spatial cues from the Hubert envelope are more consistent across frequency subbands and can clearly distinguish between spatially distributed speakers, compared to spatial cues estimated from the recorded speech or residual signal.

## Disciplines

Physical Sciences and Mathematics

## Publication Details

Cheng, E., Burnett, I. S. & Ritz, C. H. (2006). Using spatial audio cues from speech excitation for meeting speech segmentation. 8th International Conference on Signal Processing: ICSP2006 Proceedings (pp. 3067-3070). USA: IEEE Press.

# Using Spatial Audio Cues from Speech Excitation for Meeting Speech Segmentation

Eva Cheng[1], Ian Burnett, Christian Ritz
*Whisper Labs, School of Electrical, Computer, and Telecommunications Engineering*
*University of Wollongong, Wollongong, NSW, Australia 2522*
*[ecc04, ianb, critz]@uow.edu.au*

## Abstract

Multiparty meetings generally involve stationary participants. Participant location information can thus be used to segment the recorded meeting speech into each speaker's 'turn' for meeting 'browsing'. To represent speaker location information from speech, previous research showed that the most reliable time delay estimates are extracted from the Hilbert envelope of the Linear Prediction residual signal. The authors' past work has proposed the use of spatial audio cues to represent speaker location information. This paper proposes extracting spatial audio cues from the Hilbert envelope of the speech residual for indicating changing speaker location for meeting speech segmentation. Experiments conducted on recordings of a real acoustic environment show that spatial cues from the Hilbert envelope are more consistent across frequency subbands and can clearly distinguish between spatially distributed speakers, compared to spatial cues estimated from the recorded speech or residual signal.

## 1. Introduction

Facilitating the ability to efficiently access and 'browse' meeting recordings is a focus of many meeting analysis research groups [1]. One of the fundamental ways of accessing meeting recordings is to browse on the basis of each speaker's period of participation, or 'turn'. Assuming that participants in a meeting are generally stationary, Lathoud et al. [2] introduced using speaker location information for meeting speech segmentation by speaker turn. The speaker location 'cues' were Time Delay Estimations (TDE) extracted from the multichannel recorded audio, using techniques derived from Generalized Cross Correlation (GCC-PHAT [2]) and beam-forming approaches (SRP-PHAT [2]).

These concepts were extended by the authors in [3], where spatial 'cues', based on those used in Spatial Audio Coding (SAC) [4][5], represented the speaker location information extracted from the multichannel audio. SAC spatial cues were derived using psychoacoustic principles, and hence represent the *perceptual* spatial location of the sound sources [4][5]. Research in [3] showed that subband derived spatial cues detected multiple concurrent speakers in a given frame, whereas GCC-based techniques only detected the strongest speaker.

Research in [6] presented enhancements to the GCC-based algorithms for TDE from speech signals. Rather than calculating TDE from the speech signals, TDE were extracted from the Hilbert envelope of the speech residual signal obtained using Linear Prediction (LP) analysis [6]. The advantage of this approach was preservation of the pitch information whilst removing room reverberation effects during the LP analysis. With only the glottal pulses present in the residual signal, taking the Hilbert envelope removed phase ambiguities in the residual [6]. Experimental results in [6] have shown that performing cross-correlation on the Hilbert envelope results in more reliable TDE over estimates on the speech or residual signal.

This current paper combines the LP-based technique of [6] and the present authors' work in [3]. The approach proposed in this paper exploits knowledge that the meeting participants are stationary, and that speech is the primary audio source in a meeting. This results in spatial cues, as employed in [3], being calculated on the Hilbert envelope of the LP residual signal.

In the foregoing, Section 2 details the proposed approach, Section 3 describes the meeting recordings, and the corresponding results are discussed in Section 4. Section 5 concludes this paper.

## 2. Proposed System

An overview of the approach proposed in this paper is shown in Fig. 1. This paper extracts spatial cues based on the Spatial Audio Coding (SAC) analysis process [4][5]. However, rather than extracting cues directly from the recorded signals as in SAC, the approach from [6] is adopted and this paper proposes extracting spatial cues from the Hilbert envelope of the LP residual. The spatial audio cues are thus valid for speech: the phase-based spatial cues from SAC represent similar information to the TDE but in the frequency domain. Hence, this paper investigates whether performance improvements found with using the Hilbert envelope for TDE in [6] apply to the spatial cues, and in particular, the phase-based cues.

[1] Partially supported by the CSIRO ICT Centre.

**0-7803-9737-1/06/$20.00 ©2006 IEEE**

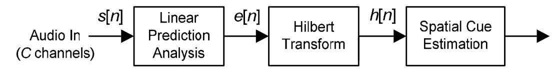


**Fig. 1.** Proposed approach

## 2.1. Linear Prediction Analysis

Linear Prediction (LP) is a technique widely employed in speech analysis. Samples in the speech signal are predicted as a weighted sum of the past $P$ samples, where $P$ is the predictor order. The residual or excitation signal ($e[n]$), is defined as the difference between the original ($s[n]$) and predicted ($\hat{s}[n]$) speech signal. The LP analysis procedure is represented as [6]:

$$\hat{s}[n] = -\sum_{k=1}^{P} a_k s[n-k]\,; \qquad e[n] = s[n] - \hat{s}[n] \qquad (1)$$

The summing weights, $a_k$, known as Linear Predictor Coefficients (LPC), were calculated using the Levinson-Durbin recursion algorithm in this paper.

The Hilbert envelope, $h[n]$, is then calculated from the residual signal, $e[n]$, and the Hilbert transform of $e[n]$ ($e_h[n]$), according to [6]:

$$h[n] = \sqrt{e^2[n] + e_h^2[n]} \qquad (2)$$

Consistent with LP speech coding techniques, this paper uses 8kHz speech and employed a 10th order LP predictor on differenced speech [7]. Differenced speech was windowed every 25ms with 10ms shift between adjacent windows. The LPC were calculated for every window, and applied to calculate the residual and Hilbert envelope. However, it is known that frame durations longer than 200-300ms can be required for reliable TDE [8]. Thus, 256ms aggregate 'superframes' for spatial cue estimation were formed from the speech, LP residual and Hilbert envelope, with 50% overlap between adjacent superframes.

## 2.2. Spatial Cue Extraction

Spatial Audio Coding (SAC) includes schemes such as Binaural Cue Coding (BCC) [4] and Parametric Stereo Coding (PSC) [5]. SAC techniques capture the perceptual spatial image of multichannel audio by extracting Interchannel Level and Time or Phase Difference cues (ICLD and ICTD/IPD) during analysis.

In this paper, the spatial cues are extracted using the SAC encoding process of [4][5]. Spectra $X_{c,m}$ for each 'superframe', $m$, are calculated using an $N$-point DFT. $X_c$ (where $m$ is omitted for simplicity) are then decomposed into $B$ frequency subbands with bandwidths matching the critical bands of human hearing [4]. The DFT coefficients in each subband, $b$, are denoted by $n \in \{A_{b-1}, A_{b-1} + 1, \cdots, A_b - 1\}$, where $A_b$ are the subband boundaries with $A_0 = 0$.

The spatial cues are calculated for each channel pair, $p$, in each subband, $b$. Mathematically, the ICLD cue is extracted according to [4]:

$$ICLD_p[b] = 10\log_{10}\left(\frac{P_2[b]}{P_1[b]}\right); P_c[b] = \sum_{k=A_{b-1}}^{A_b-1} \left|X_c[k]\right|^2 \quad (3)$$

The ICTD cue proposed in BCC did not exhibit consistent statistical trends necessary for meeting speech segmentation [3]. Thus, this paper only investigates the ICLD and the Interchannel Phase Difference (IPD) cue used in PSC, which weights the DFT bins according to magnitude. The IPD cue is obtained in subbands up to 2kHz according to [5]:

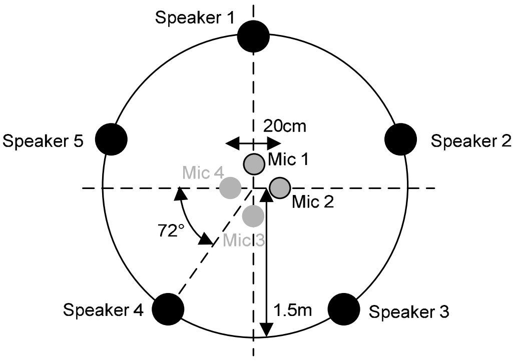


**Fig. 2.** Meeting room setup

$$IPD_p[b] = \angle\left(\sum_{k=A_{b-1}}^{A_b-1} X_1[k] X_2^*[k]\right) \qquad (4)$$

## 3. Meeting Recordings

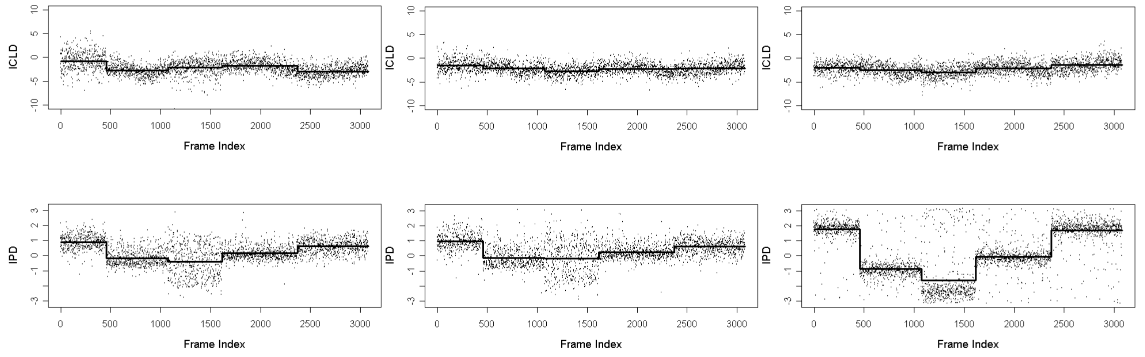
To validate the application of spatial cues to meeting audio segmentation, five loudspeakers equally spaced in a circle of 3m in diameter simulated active meeting participants in a real acoustic environment. Illustrated in Fig. 2, the recording setup used an existing spatialised audio playback system, the Configurable Hemisphere Environment for Spatialised Sound (CHESS) [9].

Clean, normalized speech from the Australian National Database of Spoken Languages (ANDOSL) simulated speech from meeting participants. Upsampled from 20kHz to 44.1kHz, silence was removed from the speech as such segments produce ambiguous spatial cues since no speaker location information exists [3].

The speech was recorded at 44.1kHz using four RODE NT2A multi-pattern microphones equally spaced in a 20cm diameter circle, as shown in Fig. 2. To simulate a meeting, each of the loudspeakers was played in turn (from Speaker 1 to Speaker 5 in Fig. 1). Each of the five speaker turns totaled 1-1.5 minutes in duration, giving a 'meeting' of about 6.5 minutes of non-overlapped speech in length. The 'meeting' recordings were downsampled to 8kHz and stored at 24 bits/sample.

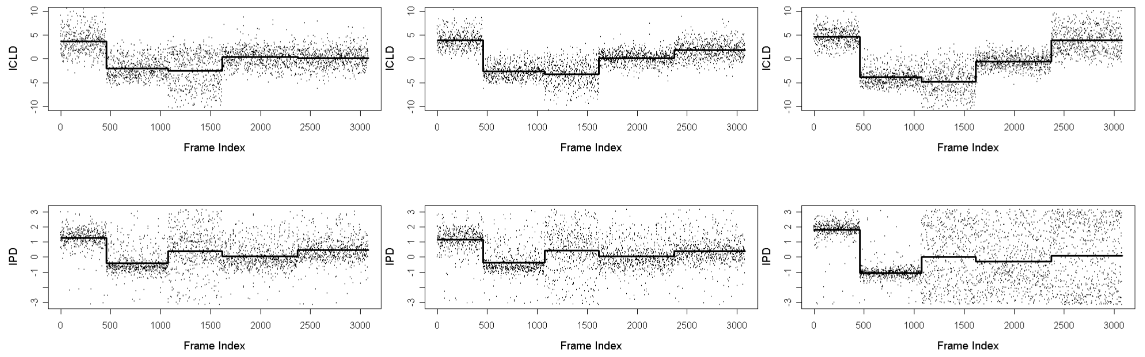
## 4. Results and Discussion

The experiments in this paper used two of the four available microphones. Mics 1 and 2 (see Fig. 2) were configured as either an omnidirectional or cardioid pattern pair of microphones. Figs. 3 and 4 show the ICLD and IPD spatial cues as a function of time for the omnidirectional and cardioid recordings, respectively. Cues were extracted from the subband centered at 928Hz, which is a frequency region where significant speech activity exists. In Figs. 3 and 4, the mean of each speaker's 'turn' is shown as the solid line. The mean was calculated according to the ground truth speaker segmentation from the original speech. The cardioid



(a) Original recordings (b) LP residual (c) Hilbert envelope

**Fig. 3.** Spatial cues from omnidirectional recordings for the subband centered at 928Hz



(a) Original recordings (b) LP residual (c) Hilbert envelope

**Fig. 4.** Spatial cues from cardioid recordings for the subband centered at 928Hz

recordings in Fig. 4 show clearer distinctions between the five speakers for the ICLD cue compared to the ICLD cues in Fig. 3 (especially Fig. 4c). In contrast, the omnidirectional recordings in Fig. 3 are better suited to the IPD cue compared to the IPD in Fig. 4 (especially Fig. 3c). These results for the ICLD and IPD are consistent with [10], which found that omnidirectional patterns were best suited for phase-based cues and cardioid for level-based cues when the spatial cues were directly estimated from the recorded speech. Thus, the results in Figs. 3 and 4 show that the influence of the microphone pattern has the same effect on the spatial cues regardless if they are extracted directly from the recorded speech, the LP residual or its Hilbert envelope.

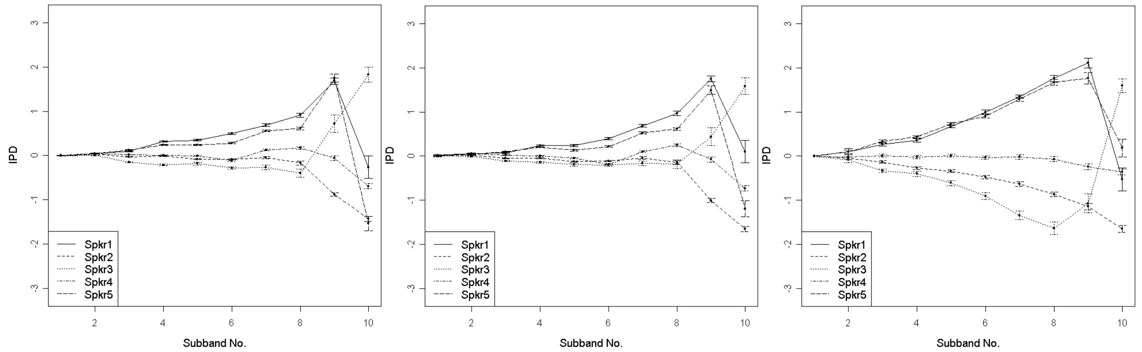
In Fig. 4, the directive cardioid pattern can reduce the effects of room reverberation and thus minimize corruption of the ICLD cue. The cardioid IPD cue shows greater cue changes between speakers 1 and 2, compared to the other speakers (especially in Fig. 4c). This is due the microphone positioning and main lobe orientation relative to the active speakers (see Fig. 2). The IPD cue is a frequency domain cross-correlation calculation (see Eq. 4) and thus requires microphones whose recorded signals differ only by environmental degradations e.g. omnidirectional pattern.

Regardless of microphone pattern, the spatial cues extracted from the Hilbert envelope in Figs. 3c and 4c show less outliers and thus improved cue ‘clusters’ are exhibited for each speaker compared to Figs. 3a-b and 4a-b. The cue clusters corresponding to the five speakers are particularly evident for the IPD in Fig. 3c. The spatial cues from the LP residual in Figs. 3b and 4b only show slight improvements over cues from the speech signal (Figs. 3a and 4a).

Figs. 5 and 6 show the mean and 95% confidence interval of the omnidirectional IPD and cardioid ICLD, respectively. The mean was calculated across all the frames from each speaker’s ‘turn’ for each subband.

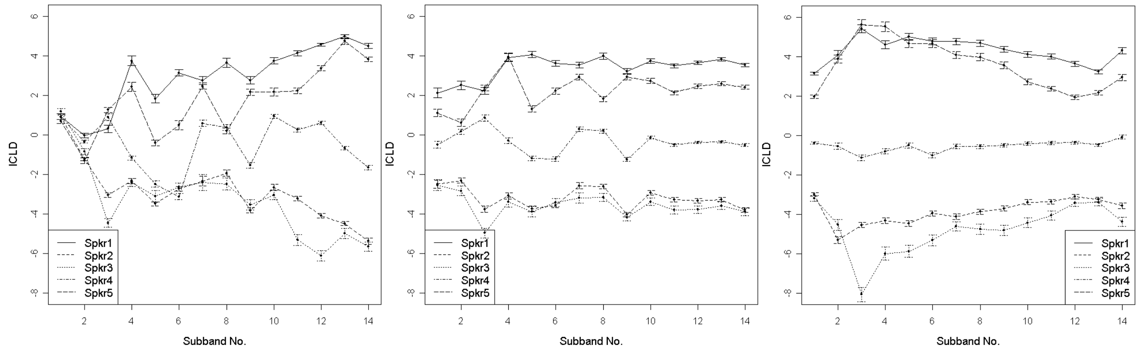
For the IPD, the omnidirectional microphones perform most consistently with the Hilbert envelope (Fig. 5c), where four of the speakers can be clearly identified in all subbands. Little improvement in the IPD is shown between the speech (Fig. 5a) and LP residual (Fig. 5b). These results show that the findings in [6] also apply to the IPD spatial cue. That is, although the pitch information dominates in the residual signal, phase ambiguities that corrupt the TDE also affect the IPD. The Hilbert envelope removes these phase ambiguities and hence performance improves with the IPD (Fig. 5c).

For the ICLD calculation, the Hilbert envelope (Fig. 6c) provides the most consistency between subbands and greater differences between cues for the five speakers. The LP residual (Fig. 6b) itself improves slightly upon the speech signal (Fig. 6a). In Figs. 6b and 6c, the consistency of the ICLD cue across the subbands for each speaker confirms that the speech residual (and hence Hilbert envelope) are noise-like spectrally flat signals. Thus, although the ICLD cues from the LP residual and Hilbert envelope do not strictly represent spatial information, the trends in Fig. 6 show that LP analysis does not remove all the speaker dependent spectral information.



(a) Original recordings (b) LP residual (c) Hilbert envelope

**Fig. 5.** Mean IPD values from omnidirectional recordings



(a) Original recordings (b) LP residual (c) Hilbert envelope

**Fig. 6.** Mean ICLD values from cardioid recordings

## 5. Conclusion

Experiments conducted in this paper on real recordings of simulated meetings have shown that spatial cues estimated from the Hilbert envelope of the Linear Prediction residual exhibit the most consistent trends across all frequency subbands, compared to cues from the recorded speech or residual signal. Spatial cues from the Hilbert envelope also showed greatest cue changes between speakers which corresponded to changing speaker location. Experiments also indicated that for spatial cues estimated directly from the recorded speech, the residual or its Hilbert envelope, omnidirectional microphone recordings exhibit the most reliable phase-based spatial cues while cardioid microphone recordings exhibit the most consistent level-based cues.

## References

[1] S. Tucker and S. Whittaker, “Accessing Multimodal Meeting Data: Systems, Problems and Possibilities,” *in Lecture Notes in Computer Science*, vol. 3361, pp. 1-11, Springer-Verlag, Berlin, 2005.

[2] G. Lathoud, I. A. McCowan, and D. Moore, “Segmenting Multiple Concurrent Speakers using Microphone Arrays,” in proc. *Eurospeech ’03*, pp. 2889-2892, Geneva, Sept. 2003.

[3] E. Cheng, et al., “Using Spatial Cues for Meeting Speech Segmentation,” in proc. *ICME ’05*, Amsterdam, July 2005.

[4] C. Faller and F. Baumgarte, “Binaural Cue Coding – Part II: Schemes and Applications,” *IEEE Trans. On Speech and Audio Processing*, vol. 11, no. 6, pp. 520-531, Nov. 2003.

[5] J. Breebaart, et al., “High Quality Parametric Spatial Audio Coding at Low Bitrates,” presented at the AES 116th Convention, Berlin, May 2004.

[6] V. C. Raykar, et al., “Speaker Localization Using Excitation Source Information in Speech,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 751-761, Sept. 2005.

[7] B. Yegnanarayana, “Enhancement of Reverberant Speech using LP Residual Signal,” *IEEE Trans. On Speech and Audio Processing*, vol. 8, no. 3, pp. 267-281, May 2000.

[8] J. DiBiase, H. Silverman, and M. Brandstein, “Robust localization in reverberant rooms,” in Microphone Arrays: Signal Processing Techniques and Applications, M. Brandstein and D. Ward, Eds., Springer-Verlag, Berlin, 2001, pp. 157-180.

[9] G. Schiemer, et al., “Configurable Hemisphere Environment for Spatialised Sound,” in proc. *ACMC ’04*, Wellington, 2004.

[10] E. Cheng, I. Burnett, C. Ritz, “Investigating Spatial Audio Coding Cues for Meeting Audio Segmentation,” presented at *AES 120$^{th}$ Convention*, Paris, May 2006.