2006

Varying microphone patterns for meeting speech segmentation using spatial audio cues

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

Varying microphone patterns for meeting speech segmentation using spatial audio cues

Abstract

Meetings, common to many business environments, generally involve stationary participants. Thus, participant location information can be used to segment meeting speech recordings into each speaker's 'turn'. The authors' previous work proposed the use of spatial audio cues to represent the speaker locations. This paper studies the validity of using spatial audio cues for meeting speech segmentation by investigating the effect of varying microphone pattern on the spatial cues. Experiments conducted on recordings of a real acoustic environment indicate that the relationship between speaker location and spatial audio cues strongly depends on the microphone pattern.

Disciplines

Physical Sciences and Mathematics

Publication Details

Cheng, E., Burnett, I. S. & Ritz, C. H. (2006). Varying microphone patterns for meeting speech segmentation using spatial audio cues. Q. He, Y. Rui, S. Yang & Y. Zhuang In Pacific-Rim Conference on Multimedia, Nov 2-4, Hangzhou, China. LNCS - Advances in Multimedia Information Processing (PCM 2006), 4261 221-228.

Varying Microphone Patterns for Meeting Speech Segmentation using Spatial Audio Cues

Eva Cheng[1,2], Ian Burnett[1], Christian Ritz[1]

[1] Whisper Labs, School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Wollongong NSW Australia 2522
Ph: +61 (0)2 4221 3785 Fax: +61 (0)2 4221 3236
{ecc04, ianb, critz}@uow.edu.au

Abstract. Meetings, common to many business environments, generally involve stationary participants. Thus, participant location information can be used to segment meeting speech recordings into each speaker's 'turn'. The authors' previous work proposed the use of spatial audio cues to represent the speaker locations. This paper studies the validity of using spatial audio cues for meeting speech segmentation by investigating the effect of varying microphone pattern on the spatial cues. Experiments conducted on recordings of a real acoustic environment indicate that the relationship between speaker location and spatial audio cues strongly depends on the microphone pattern.

Keywords: spatial audio cues, meeting audio analysis, microphone arrays.

1 Introduction

Meetings recordings are currently difficult to access offline as users potentially have to search through hours of audio/video data to find segments of interest. Efficient 'browsing' of meeting recordings is thus of great interest to many business environments. Research groups currently focus on meeting analysis to facilitate effective meeting browsing [1]. Browsing requires the meeting to be segmented and annotated with semantically meaningful information such as speaker identification, speaker location, speech summary, transcript or level of participant interaction e.g. monologue, group discussion or presentation.

Many of the annotations can be derived from the meeting audio recordings alone. A fundamental way to segment meeting audio is by each speaker's period of participation, or 'turn'. For such segmentation, Lathoud et al. represented speaker location information as Time Delay Estimations (TDE) derived from omnidrectional recordings, using Generalized Cross Correlation (GCC) based techniques such as GCC-PHAT and SRP-PHAT [2].

The authors' previous work extended the concept of using speaker location information for meeting speech segmentation by representing speaker location information with spatial audio cues [3]. The cues are derived from the spatial cues in

[2] Partially supported by the CSIRO ICT Centre, Australia.

Spatial Audio Coding (SAC) [4, 5]. When applied to meeting speech segmentation, previous work found that the spatial cues detected multiple concurrent speakers, whereas GCC-based approaches tended to only detect the strongest speaker in a given speech frame [3].

This paper further investigates the validity of using spatial audio cues for meeting speech segmentation. Simulated meetings recorded with different microphone patterns are analyzed to explore the how microphone pattern affects the relationship between the spatial cues and speaker location. SAC techniques have not been previously investigated with microphone array recordings. Rather, research has focused on resynthesising spatialised recordings e.g. 5.1 surround [6].

In the remainder of this paper, Section 2 outlines the spatial cues extraction from the recorded speech. Section 3 describes the meeting recording setup, while Section 4 presents the results obtained from these recordings. Section 5 concludes the paper.

2 Spatial Audio Coding

SAC aims to compactly represent multichannel audio for storage and transmission over mediums such as the Internet. In the encoder, spatial cues are extracted from the *C*-channel input audio (where $C > 1$) and the audio is downmixed into *D* channels (where $D < C$). The *D*-channel downmix and associated ‘side information’ (spatial cues) are sent to the decoder for resynthesis into *C*-channels which aim to recreate the original perceptual ‘spatial image’ for the user.

This paper explores the use of spatial audio cues for the purposes of meeting speech segmentation by speaker turn. The spatial audio cues, derived from SAC, represent the *perceptual* speaker location information contained in the speech recordings. Psychoacoustic studies have shown that humans localize sound sources with two main cues: Interaural Level Difference (ILD) and Interaural Time Difference (ITD) [7]. These psychoacoustic concepts are adopted by SAC approaches to derive level and time or phase-based spatial cues from multichannel audio. This paper employs spatial cues originally introduced by the following SAC schemes: Binaural Cue Coding (BCC) [4] and Parametric Stereo Coding (PSC) [5].

2.1 Spatial Audio Coding Cues

BCC and PSC encoders accept *C*-channel input, where $C = 2$ for PSC. Each time-domain channel, *c*, is split into *M* frames using 50% overlapped windows. The frequency-domain spectrum, $X_{c,m}[k]$, is obtained by an *N*-point Discrete Fourier Transform (DFT) for each channel, *c*, and frame, *m* [7][8]. The human hearing system uses non-uniform frequency subbands known as ‘critical bands’ [7]. Thus, $X_{c,m}[k]$ is decomposed into non-overlapping subbands of bandwidths that match these critical bands. The DFT coefficients in each subband are denoted by $k \in \{A_{b-1}, A_{b-1}+1, \ldots, A_b - 1\}$, where A_b are the subband boundaries and $A_0 = 0$.

BCC calculates the following spatial cues between each input channel c ($2 \leq c \leq C$) and a reference channel (taken to be channel one i.e. C = 1), for each frame, *m*, and each subband, *b* [4]:

- Inter-Channel Level Difference (ICLD)

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

- Inter-Channel Time Difference (ICTD), which estimates the average phase delay for subbands below 1.5kHz. Subbands above 1.5kHz estimate the group delay.

The PSC encoder extracts the following spatial cues between the two input channels for each frame, *m*, and each subband, *b* [5]:

- Inter-channel Intensity Difference (IID)

$$IID[b] = 10\log_{10}\left(\frac{\sum_{k=A_{b-1}}^{A_b - 1} X_1[k]X_1^*[k]}{\sum_{k=A_{b-1}}^{A_b - 1} X_2[k]X_2^*[k]}\right) , \tag{2}$$

- Inter-channel Phase Difference (IPD), limited to the range $-\pi \leq IPD[b] \leq \pi$,

$$IPD[b] = \angle\left(\sum_{k=A_{b-1}}^{A_b - 1} X_1[k]X_2^*[k]\right) . \tag{3}$$

The human auditory system is less sensitive to interaural phase differences at frequencies greater than approximately 2kHz [7]. Thus, BCC splits the ICTD calculation at 1.5kHz and PSC only estimates the IPD cue for subbands below 2kHz.

Previous research showed that the ICTD from BCC did not exhibit a strong nor consistent relationship with speaker location [3]. Thus, this paper implements a combined BCC and PSC encoder where only the ICLD (which is effectively the same calculation as the IID) and IPD cues are estimated.

3 Meeting Recordings

To simulate a real meeting environment, recordings were made in an immersive spatial audio playback system. Fig. 1 illustrates the recording setup. The Configurable Hemisphere Environment for Spatialised Sound (CHESS) [8] was used to simulate a meeting with five participants equally spaced (i.e. 72° apart) in a circle approximately 3m in diameter. Each participant was represented by a loudspeaker which played clean speech sourced from one person.

Clean speech was obtained from the Australian National Database of Spoken Languages (ANDOSL). Five native Australian speakers, two female and three male, were chosen as the meeting 'participants'. The ANDOSL speech files were upsampled to 44.1kHz from 20kHz, and normalized with silence removed. Previous work showed that silence segments produced ambiguous spatial cues, since no speaker location information exists in such segments [3].

To simulate a meeting, each of the five loudspeakers was played in turn (from Spkr1 to Spkr5 in Fig. 1). Each participant's turn ranged from 1-1.5 minutes in duration. This resulted in a 6.5 minute long 'meeting' of non-overlapped speech.

To record the speech, two AKG C414 B-XL II multi-pattern microphones were placed in the centre of the 'meeting environment', spaced 20cm apart. All experiments utilized the two microphones in the Mic 1 and Mic 2 positions, as shown in Fig. 1. Microphones recorded the speech which was then sampled at 44.1kHz and stored at 24 bits/sample.

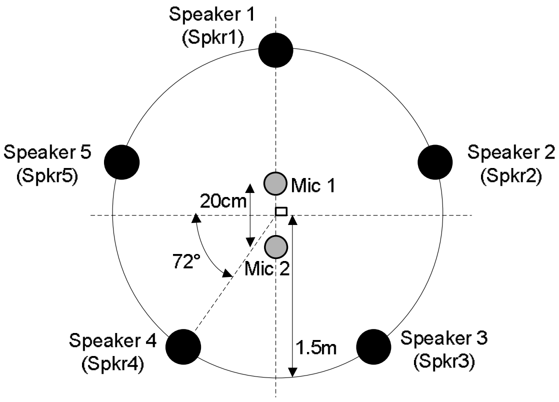


Fig. 1. Recording setup

4 Results

The simulated 'meetings' were recorded in CHESS [8] using the two microphones configured as a pair of omnidirectional, cardioid, hypercardioid, or figure-8 pattern microphones. For each of the four microphone patterns under study, spatial cues were estimated from the pair of recordings using the combined BCC/PSC spatial audio encoder. At a sampling rate of 44.1kHz, this resulted in a decomposition of 21 subbands and a DFT of length 2048 was used.

Fig. 2 plots the ICLD and IPD spatial cues as a function of time. The mean of each speaker's 'turn' is shown as the solid line. The mean was calculated based upon the ground truth segmentation from the original speech. The shown ICLD is taken from the subband centered at 2.5kHz, which is a frequency region that exhibits strong speech activity. In contrast, the displayed IPD is taken from the subband centered at 382Hz, since pitch information dominates at low frequencies. Fig. 2 shows that the spatial cues vary significantly depending on the microphone pattern used.

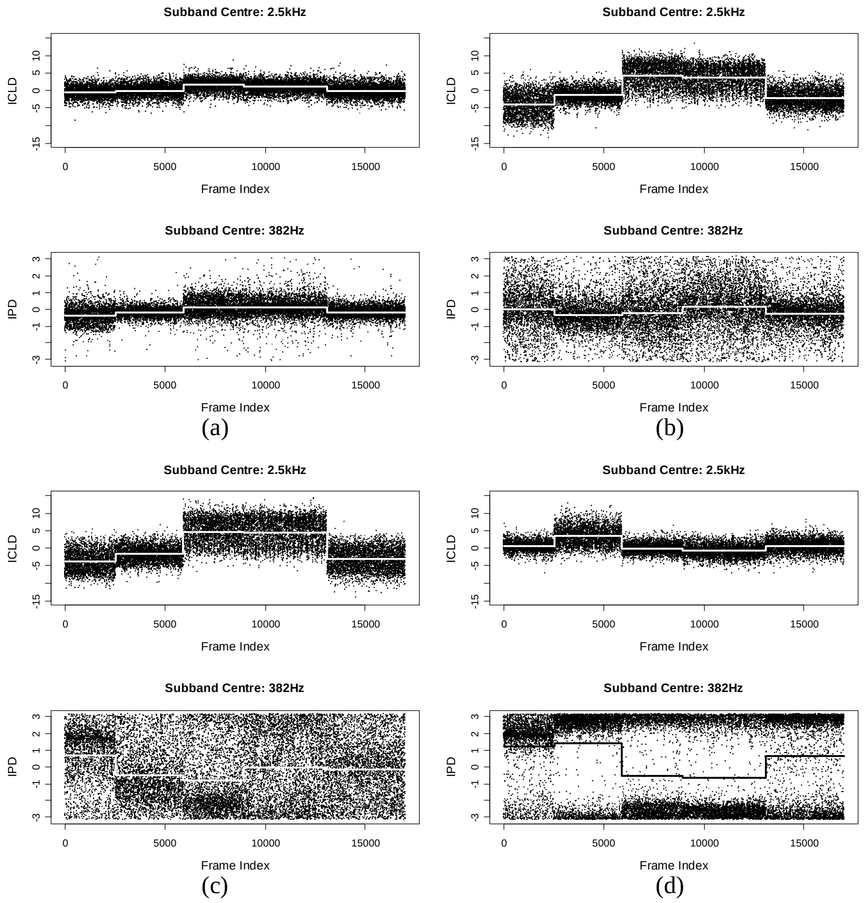
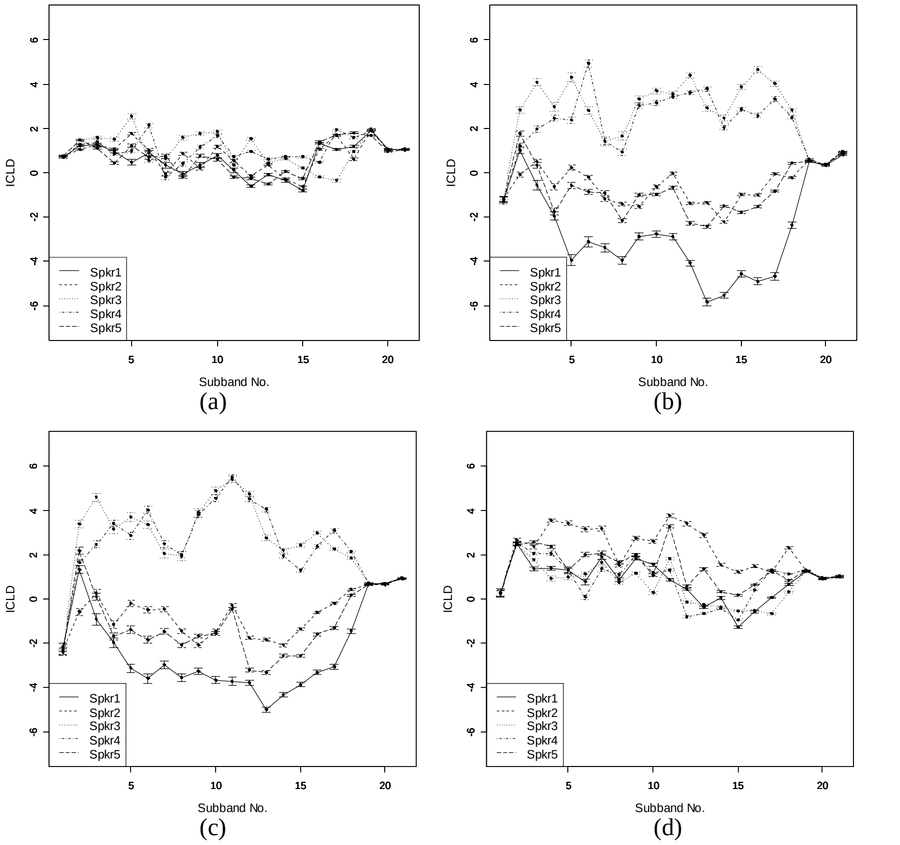


Fig. 2. ICLD and IPD as a function of time: (a) omnidirectional, (b) cardioid, (c) hypercardioid, and (d) figure-8 microphones

4.1 ICLD

Fig. 3 illustrates the mean and the 95% confidence interval (CI) of the ICLD cue in each subband for each speaker. The mean was calculated across all the frames from each speaker's 'turn' for each subband.

The cardioid (Fig. 3b) and hypercardioid (Fig. 3c) patterns clearly show three groups of ICLD trends across the subbands. Ideally, the five different speakers should exhibit five distinct ICLD trends. However, this is not possible with one microphone pair: in that case, sound localization is limited to sources that are not equidistant between the two microphones. Thus, the three trends seen in Figs. 3b and 3c correspond to Speaker 1, and the equidistant pairs of Speakers 2 and 5, and Speakers 3 and 4 (see Fig. 2). In contrast, the omnidirectional (Fig. 3a) and figure-8 (Fig. 3d)



(a) (b) (c) (d)

Fig. 3. Mean ICLD values: (a) omnidirectional, (b) cardioid, (c) hypercardioid, and (d) figure-8 recordings cannot clearly distinguish between the five spatially separated speakers based on the ICLD cue.

The authors’ previous work showed that the ICLD represented spatial information, independent of the speaker characteristics [3]. These findings are confirmed by the cardioid and hypercardioid results in Figs. 3b and 3c. These microphone patterns are better suited to the level-based spatial cue because the single lobe directionality limits the influence of background noise that can corrupt ICLD estimation.

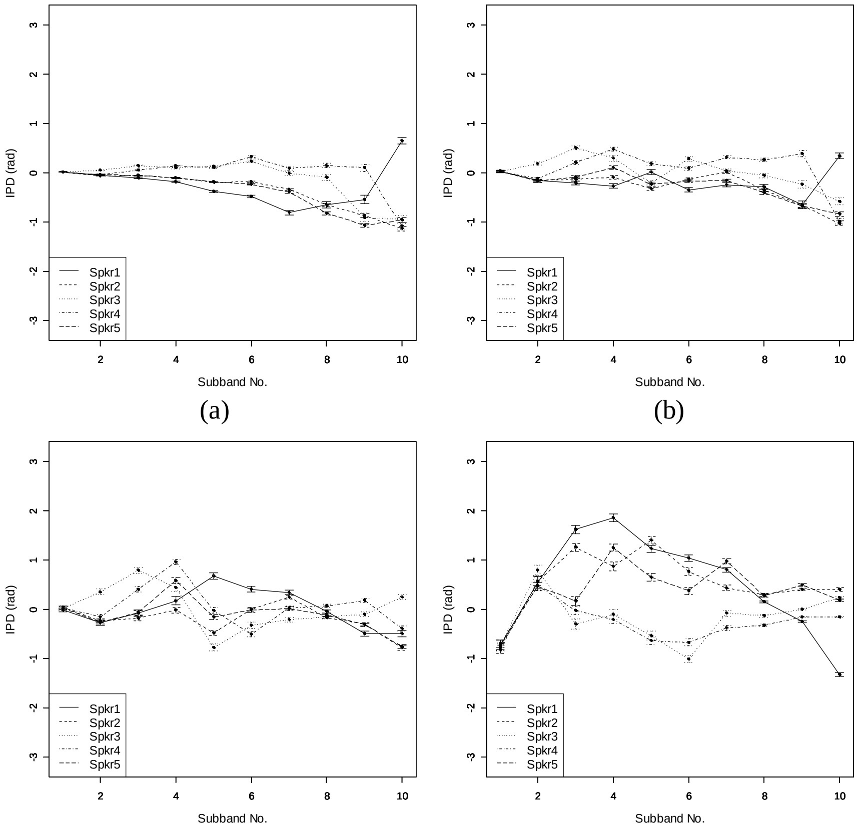
In Fig. 3, the ICLD cue does not show inter-speaker trends for very low or very high frequencies. For the first (centered at 27Hz) and last three subbands (centered at 12kHz, 15kHz, 19kHz), all five speakers exhibit similar means for all microphone patterns. At high frequencies, although the recordings were sampled at 44.1kHz, the original speech was sampled at 20kHz and hence no frequency information exists above 10kHz. In the low frequency subband, however, there is little speech activity in this frequency region and thus minimal spatial information exists.

In Fig. 3, the mean ICLD value per speaker varies across the subbands. This is due to the prominence of speech activity around certain frequency regions. In addition, the ICLD calculation already combines the contribution from a range of DFT frequency bins (see Equation 1).

4.2 IPD

Fig. 4 illustrates the mean and 95% CI of the IPD cue in each subband for each speaker's 'turn'. The opposite trends to Fig. 3 are shown in Fig. 4. In Fig. 4, the IPD cue from the cardioid (Fig. 4b) and hypercardioid (Fig. 4c) pattern recordings do not show significant differences between the five speakers. Similarly to the ICLD cue, Speakers 2 and 5, and Speakers 3 and 4 should exhibit similar IPD cues due to the microphone placement. The pattern that best shows this trend is the omnidirectional microphone (Fig. 4a). For the figure-8 recordings (Fig. 4d), the cues from Speakers 3 and 4 match more clearly than those from Speakers 2 and 5.

The superior performance of the omnidirectional microphones for IPD cue estimation is consistent with previous work, which found that omnidirectional patterns are best suited to TDE [2]. The TDE calculation, like the IPD cue (see Equation 3), involves cross-correlation estimation. Such calculations require spatially distributed microphones that record the same signal but vary according to degradation from the acoustic environment. The omnidirectional pattern fits this requirement, while the directional patterns record signals that vary depending on the source



(a)

(b)

(c)

(d)

Fig. 4. Mean IPD values: (a) omnidirectional, (b) cardioid, (c) hypercardioid, and (d) figure-8

direction relative to the main pickup lobe/s. Thus, the figure-8 pattern performs better than the cardioid and hypercardioid patterns for the IPD cue in Fig. 4. By having two main pickup lobes, figure-8 patterns can capture signals which do not differ as much between spatially distributed microphones compared to single main lobe patterns.

Similarly to the ICLD cue, the means of the IPD vary across the subbands. In addition, the means of the IPD from the different speakers converge to similar values in the first subband (centered at 27Hz). The lack of spatial information in the first subband is because of little pitch or speech activity in these frequency regions. Due to the psychoacoustically motivated subband decomposition, low frequency subbands also contain fewer DFT bins. Hence, frequency bins that do not contain information are more likely to corrupt the spatial cue estimation in these smaller subbands.

5. Conclusion

Experiments in this paper have shown that spatial audio cues, derived from spatial audio coding, do strongly correspond to changing speaker location. Thus, spatial cues are valid for segmenting meeting recordings into speaker ‘turns’. Recordings made in a real acoustic environment simulating a meeting showed that the microphone pattern significantly affected the spatial cue trends. Experiments showed that directional microphone patterns such as cardioid and hypercardioid were best suited to level-based cues. In contrast, the omnidirectional pattern exhibited the most consistent trends for phase-based cues. Thus, appropriate microphone pattern choice can help to reduce spatial cue degradation from room reverberation and background noise, without requiring post-processing of the recordings, spatial cues, or modification of spatial cue estimation techniques. The experimental results in this paper also suggest that spatial audio coding techniques are suitable for coding microphone array signals.

References

1. Tucker, S. and Whittaker S.: Accessing Multimodal Meeting Data: Systems, Problems and Possibilities. Lecture Notes in Computer Science, Springer-Verlag, vol. 3361, (2005) 1-11.
2. Lathoud, G., McCowan, I., and Moore, D.: Segmenting Multiple Concurrent Speakers using Microphone Arrays. Proc. Eurospeech ’03, Geneva (2003) 2889-2892.
3. Cheng, E. et al.: Using Spatial Cues for Meeting Speech Segmentation. Proc. ICME ’05, Amsterdam, (2005).
4. Faller, C. and Baumgarte, F.: Binaural Cue Coding – Part II: Schemes and Applications. IEEE Trans. On Speech and Audio Processing, vol. 11, no. 6, (2003) 520-531.
5. Breebaart, J. et al.: High Quality Parametric Spatial Audio Coding at Low Bitrates. AES 116th Convention, Berlin, (2004).
6. Breebaart, J. et al.: MPEG Spatial Audio Coding/MPEG Surround: Overview and Current Status. AES 119th Convention, New York, (2005).
7. Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization, MIT Press, Cambridge, (1997).
8. Schiemer, G. et al.: Configurable Hemisphere Environment for Spatialised Sound. Proc. Australian Computer Music Conference (ACMC ’04), Wellington, (2004).