THE EFFECT OF MICROPHONE DIRECTIVITY PATTERNS ON SPATIAL CUES FOR REVERBERANT MULTICHANNEL MEETING SPEECH ANALYSIS

E. Cheng[1], I. S. Burnett[2], C. Ritz[1]

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

[2]School of Electrical and Computer Engineering
Royal Melbourne Institute of Technology, Melbourne VIC Australia 3000
ian.burnett@rmit.edu.au

ABSTRACT

Multiparty meetings common to many business environments often have participants who are generally stationary. Hence, active speakers can be disambiguated by location, and meeting analysis research groups have proposed the use of speaker location information (spatial cues) for meeting segmentation and higher level analysis. As the cues are estimated from multi-microphone recordings, this paper studies the effect of varying microphone directivity patterns on the spatial cue accuracy and reliability. Results from theoretical simulations and recordings from a real reverberant environment suggest that different spatial cues (based on inter-microphone signal time delays or amplitude level differences) optimally respond to different microphone directivity patterns, where time delay accuracy was found to be independent of the relative microphone configuration.

1. INTRODUCTION

Multiparty meetings are common to many business, educational, and research environments; however, meeting audio recordings are currently not automatically segmented and annotated in a semantically meaningful manner for users to efficiently access the parts of meetings they are interested in. In structured meetings, participants are generally stationary hence their location information (spatial cues) derived from multi-microphone recordings can be used to segment and analyse the meeting according to each participant's period of interaction.

The use of speaker location information for meeting speech segmentation was proposed by Lathoud et al. [1], who introduced the use of Time Delay Estimations (TDE) calculated from multi-microphone recordings. Specifically, Lathoud et al. investigated the Generalised Cross Correlation with PHAse Transform (GCC-PHAT) technique [2], and further suggested the use of the Steered Response Power with PHAse Transform (SRP-PHAT) beamforming method to address the shortcomings of GCC-PHAT [3].

However, meetings have been traditionally recorded using multiple omnidirectional microphones, often arranged in a circular or uniform linear array and placed at the centre of the table. More recently, 'smart' meeting rooms have deployed omnidirectional or cardioid microphones in custom arrays [4] or distributed around the room (e.g., on walls) [5] for improved meeting speech capture. To extend and complement these studies and study the effect of microphone directivity patterns on spatial cue performance, the authors previously investigated deploying first-order microphone directivity patterns (e.g., omnidirectional, cardioid, hypercardioid, figure-8), finding that time and phase-based inter-microphone spatial cues most reliably responded to the omnidirectional pattern, whilst amplitude level-based cues were best suited to a cardioid directivity pattern [6]. These findings, however, were conducted using multi-pattern microphones deployed in a real reverberant recording environment, where the microphone patterns studied were limited to commercially available configurations.

This paper addresses the physical limitations of commercial microphones by simulating a range of directivity patterns to determine the 'optimal' microphone pattern for each type of spatial cue. Furthermore, microphone array configurations were also varied, with recordings in a real reverberant environment conducted to verify the simulation results.

In the remainder of this paper, Section 2 introduces the spatial cues and microphone directivity patterns studied, with the meeting simulation and recording environments detailed in Section 3. Section 4 presents and discusses the experimental results obtained, with the paper concluded in Section 5.

2. BACKGROUND

Figure 1 illustrates the meeting speech spatial analysis system studied. This paper addresses the first two stages: the multichannel recording and spatial cue extraction methods.

2.1 Spatial Cue Extraction

Spatial cues can be calculated in the time or frequency-domain, where time-domain cues such as Time-Delay Estimation (TDE) can be more computationally efficient but less effective with multiple active speakers [7]. To complement the TDE with frequency domain cues that can handle multiple speakers through exploiting sparsity of the speech spectrum, this paper extends upon the authors previous investigations into the use of spatial cues for meeting speech segmentation, with cues derived from Spatial Audio Coding (SAC) [8][9][10].

2.1.1 Time-Delay Estimation

Generalisation Cross-Correlation (GCC) is a technique commonly applied to deriving TDE from two microphone channels [2]. Mathematically, GCC is given by:

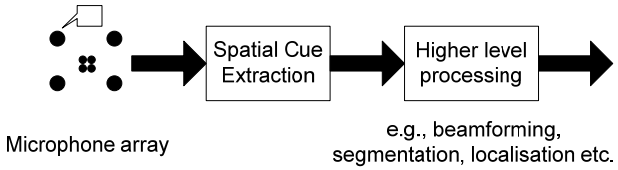


Figure 1 - Meeting spatial analysis paradigm

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

where the Discrete Fourier Transform (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:

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

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

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

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 [2]:

$$W[k] = \left| X_1[k] \cdot X_2^*[k] \right|$$

2.1.2 Spatial Audio Coding Cues

Spatial Audio Coding (SAC) aims to efficiently encode and transmit multichannel audio e.g., 5.1 surround sound, and includes schemes such as Binaural Cue Coding (BCC) [8], Parametric Stereo Coding (PSC) [9], and MPEG Surround [10]. SAC techniques capture the perceptual spatial image of multichannel audio by extracting Inter-channel Level, Time or Phase Difference, and Correlation cues (ICLD, ICTD/IPD, and ICC respectively) during the analysis stage.

This paper studies the spatial cue that is common to all the coders of [8][9][10]: the level difference cue (ICLD). The time and phase-based cues from the spatial audio coders were not investigated due to their poor performance in representing spatial information, where time/phase-cues were subsequently omitted from the MPEG Surround standard [10].

Mathematically, the spectra $X_{c,m}$ for each frame m and channel c are calculated using an N-point DFT. X_c (where m is omitted for brevity) are then decomposed into B frequency subbands with bandwidths matching the critical bands of human hearing [8]. The spatial cues are calculated for each channel pair p, in each subband b. Mathematically, the ICLD cue is extracted according to [8][9][10], where A_b are the subband boundaries with $A_0 = 1$:

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

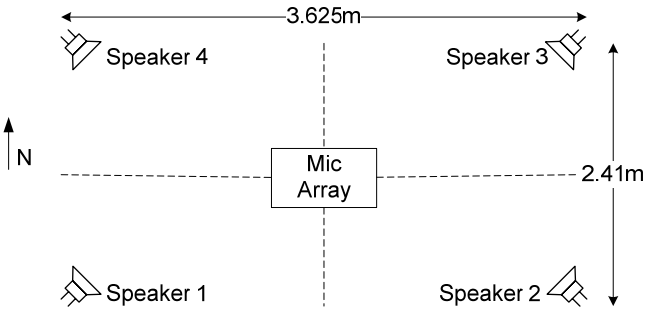
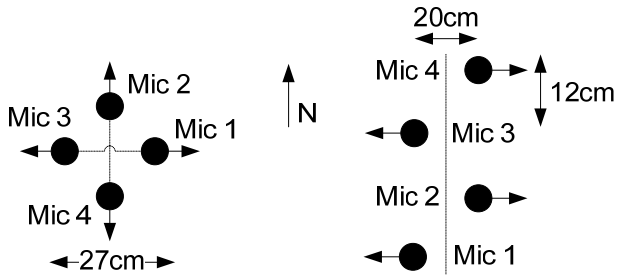


Figure 2 - Meeting recording setup



(a) Circular array (b) Offset line array

Figure 3 - Microphone array configurations

2.2 Microphone Directivity

At present, microphones are commercially available in limited directivity patterns, most commonly omnidirectional, cardioid, figure-8, hypercardioid, and superdirective e.g., shotgun. However, there have been few studies investigating which microphone directivity patterns are best suited to speech spatial analysis, and whether patterns 'in between' those commercially available are optimal for capturing speech location information.

Mathematically, the microphone directivity pattern is represented in the polar domain as [11]:

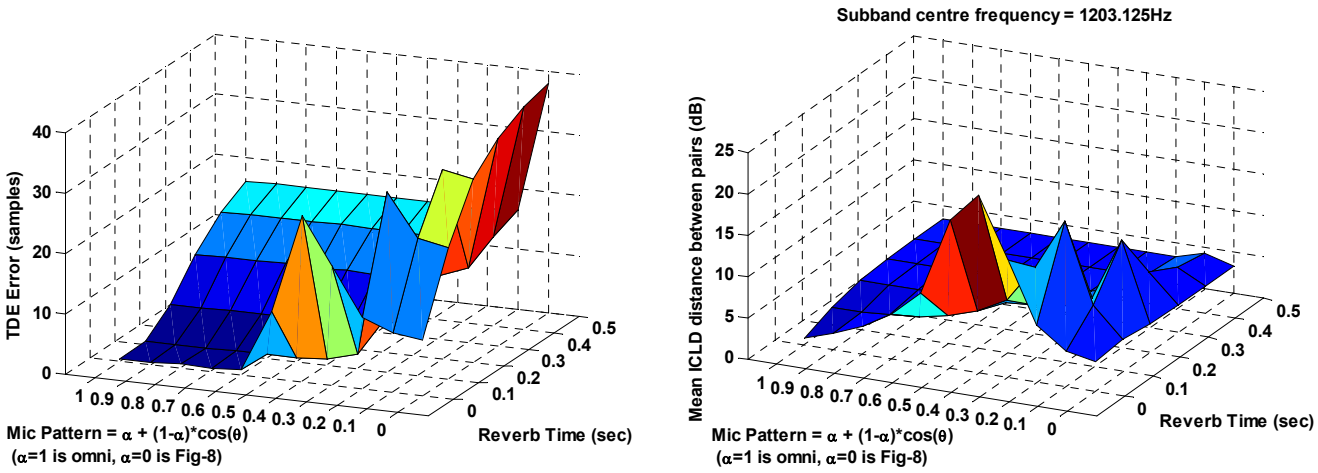
$$A(\theta) = \alpha + (1 - \alpha)\cos(\theta)$$

where θ denotes the source azimuth as measured according to convention from the positive x-axis, and $0 \le \alpha \le 1$ is the parameter controlling the directivity, where $\alpha = 1$ is omnidirectional, $\alpha = 0.5$ is cardioid, and $\alpha = 0$ is figure-8.

To study microphone directivity patterns that are not easily commercially available, varying microphone directivity can be simulated as part of the microphone response in room acoustics modelling e.g., in the Allen and Berkeley image method [12] employed in this paper.

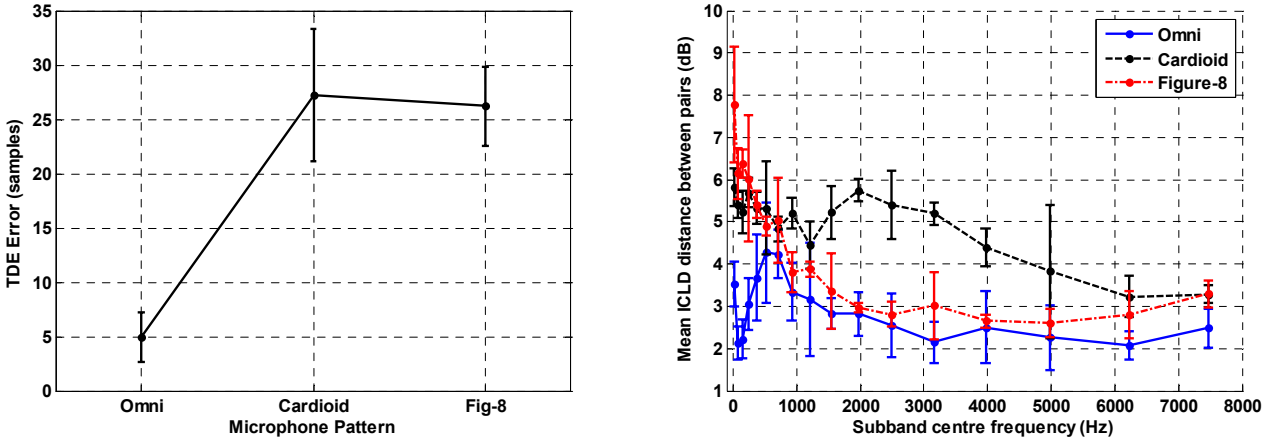
3. MEETING RECORDINGS

Figure 2 illustrates four loudspeakers placed around a 3.625m × 2.41m room simulating a 'meeting' scenario. Recorded by four microphones placed at the centre of the room, a circular and offset line microphone array were evaluated to determine the effect of varying microphone array configuration (shown in Figure 3, the arrows depict the microphone orientations). The inter-microphone spacings were chosen in accordance to typical table-top meeting recording arrays [1][3] and large enough to obtain meaningful TDEs and inter-microphone level differ-



(a) TDE RMSE (b) ICLD MSD

Figure 4 - Circular microphone array simulation results



(a) TDE RMSE (b) ICLD MSD

Figure 5 - Circular microphone array recordings results

ences. The offset line array configuration in Figure 3b was chosen to study a non-standard array configuration, where the more traditional uniform linear array returned similar results (not presented here for brevity).

The 'meeting' setup was theoretically modelled using Allen and Berkeley's image method [12], with reverberation times (RT60) varying from anechoic (RT60 = 0s) to RT60 = 0.5s. Real recordings were then conducted with four Genelec 1029A loudspeakers recorded with four Rode NT2A multipattern microphones (omnidirectional, cardioid, and figure-8). The reverberant room exhibited approx. RT60 = 300ms, which is typical of an office space.

Anechoically recorded speech from the Australian National Database of Spoken Languages (ANDOSL)[1] simulated speech from meeting participants. In turn, each loudspeaker played out one speech sentence (approx. 2s in duration), thus giving a 'meeting' of approx. 10s of non-overlapped speech in length. Recorded at 48kHz, microphone signals were downsampled to 16kHz for spatial cue extraction.

[1] http://newandosl.rsise.anu.edu.au/andosl

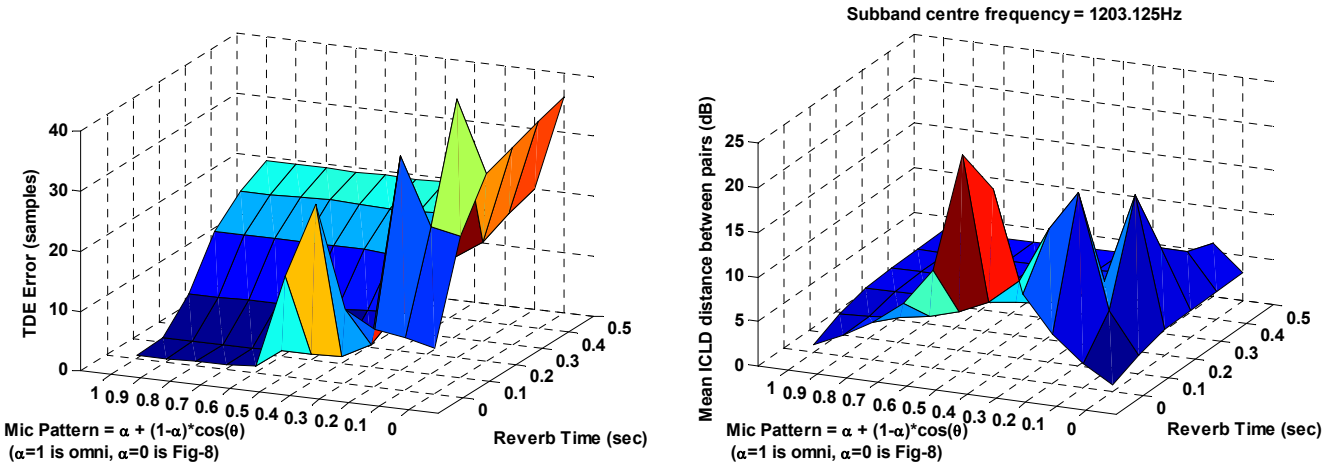
4. EXPERIMENTAL RESULTS

For all experiments, frames were 50% overlapped and 32ms in length. To evaluate the effect of varying microphone directivity pattern on spatial cue performance, the following error measures were employed, with 95% confidence intervals shown:

- TDE Root mean squared error (RMSE), as calculated between the estimated value $\hat{\tau}$ and the ground-truth τ (since the microphone and loudspeaker positions are known). In the following, N, M and P are the total number of speakers, frames and microphone pairs, respectively:

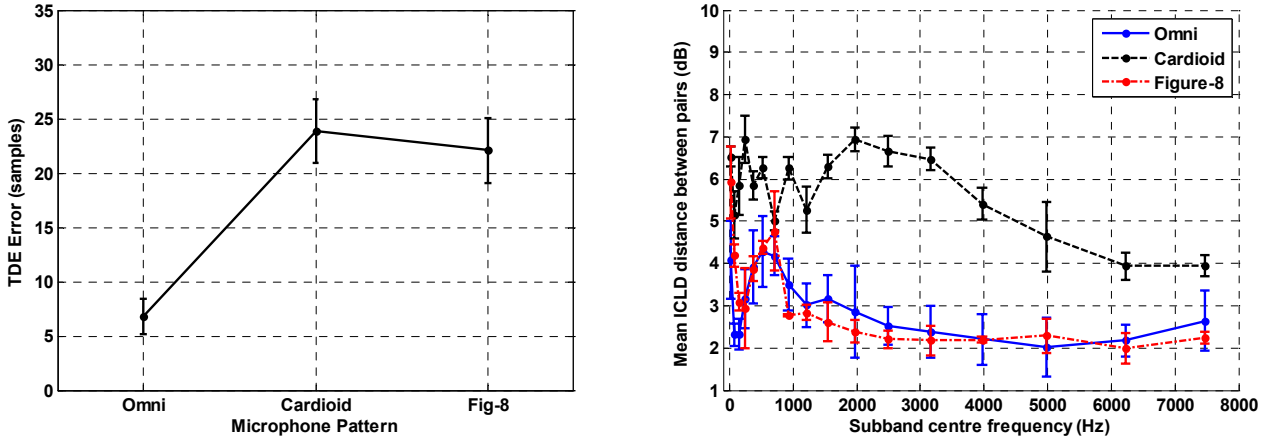
$$\text{RMSE} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{P}\sum_{p=1}^{P}\sqrt{\left(\tau_p(n,m)-\hat{\tau}_p(n,m)\right)^2}\right)\right)$$

- ICLD – Mean squared distance (MSD) between ICLD estimated between microphone pairs, where a large distance is preferred as it indicates that the pairs contain distinct information. In the following, q denotes the combinations of ICLD pairs where Q is the total number of pairs, $Q = \binom{P}{2}$:



(b) TDE RMSE (b) ICLD MSD

Figure 6 – Offset microphone line array simulation results



(b) TDE RMSE (b) ICLD MSD

Figure 7 – Offset microphone line array recordings results

$$\mathrm{MSD} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{Q}\sum_{q=1}^{Q}\sqrt{\left(ICLD_{q,1}(n,m) - ICLD_{q,2}(n,m)\right)^2}\right)\right)$$

4.1 Circular Array Microphone Configuration

Figure 4a shows the TDE RMSE from theoretical simulations, where it can clearly be seen that less directivity (i.e., higher α) yields lower TDE RMSE error, with the omnidirectional pattern exhibiting 0.4 sample error in anechoic conditions to 16.2 samples at RT60=0.5s. The RMSE is comparable for omnidirectional-type directivity patterns from $0.6 \le \alpha \le 1$, but performance significantly deteriorates with unidirectional and bidirectional patterns, where $0 \le \alpha \le 0.5$. Figure 4a clearly shows that unidirectional directivity leads to particularly erroneous TDEs at lower RT60, with the RMSE peaking at 27.5, 29.3 and 31.5 samples for α = 0.4, 0.2 and 0.1, respectively. At RT60 > 0.4 for $0.1 \le \alpha \le 0.5$, however, the RMSE becomes comparable to the omnidirectional directivity patterns. In contrast, when α = 0 (figure-8 bidirectional pattern), errors decrease at low RT60 to 9.2 samples in anechoic conditions and monotonically increase to 38.1 samples at RT60 = 0.5s. The theoretical results and trends seen in Figure 4a are consistent with experimental findings [6][7] which attribute omnidirectional microphone directivity patterns to be best suited to TDE, due to the nature of cross-correlation calculations. Further, TDE by cross-correlation methods can lead to erroneous results when the signals are dissimilar, which is the case with unidirectional patterns (in particular the cardioid, as it has no bidirectional component at all). The improved performance of unidirectional patterns at higher RT60 can be attributed to higher correlations found between microphone signals with the reverberant speech not present in anechoic conditions.

Figure 4b illustrates the ICLD MSD, where a larger distance indicates more useful spatial information between the microphone pairs. Although only the subband centred at approx. 1.2kHz is shown, similar trends were exhibited in other frequency subbands. It can be clearly seen from Figure 4b that unidirectional patterns give the greatest MSD between microphone pairs, peaking at 21.3dB at α = 0.4, which is a cardioid pattern with a minimal bidirectional component. With increased reverberation, however, a larger bidirectional component yields higher distances, with the MSD reaching 17dB (at α = 0.2, RT60 = 0.1s) and 13dB (at α = 0.1, RT60 = 0.2s). These trends seen in Figure 4b are consistent with experimental findings [6], and can be attributed to directional microphone characteristics minimising the signal amplitude corruption from reverberant speech components.

Figure 5a shows the TDE RMSE as obtained from real recordings in a room of approx. RT60 = 300ms. It can be seen that the low error performance of 4.9 samples from the omnidirectional pattern confirms the theoretical findings in Figure 4a. Moreover, the trend of the figure-8 directivity pattern slightly outperforming the cardioid with a lower RMSE, as seen in Figure 4a, is also shown with the commercial microphones.

Figure 5b depicts the ICLD MSD across the frequency subbands, and although the bidirectional figure-8 pattern exhibits high MSD up to 7.8dB at low frequencies, the cardioid outperforms the figure-8 and omnidirectional patterns by up to 3dB from approx. 1.5 – 6kHz, a frequency range where speech energy is generally strongest.

4.2 Offset Line Array Microphone Configuration

Figure 6a illustrates the TDE RMSE as obtained from the offset microphone line array in Figure 3b. Comparing Figures 4a and 6a, it can be clearly seen that similar trends are exhibited despite the varied microphone array configuration. Slightly higher TDE RMSE at high RT60 from omnidirectional directivity result from the offset line array: 19.3 samples ($\alpha = 1$, RT60 = 0.5), compared to 16.2 in Figure 4a. In contrast, the figure-8 bidirectional pattern exhibits lower RMSE in Figure 6 (35.9 samples) compared to Figure 4a (38.1 samples). And although the relative amplitude relationships between the peaks from $0.1 \le \alpha \le 0.4$ remain consistent between Figures 4a and 6a, the RMSE values vary in Figure 6a to be 12.4, 35.8, and 43 samples for $\alpha = 0.4$, 0.2 and 0.1, respectively.

Figure 6b shows the ICLD MSD from the frequency subband centred at approx. 1.2kHz, although results were similar across all subbands. The trends in Figure 4b are almost identically seen in Figure 6b, but for amplitude variations in the peaks of 24.3dB ($\alpha = 0.5$, RT60 = 0s), 19.4dB ($\alpha = 0.2$, RT60 = 0.1s), and 17.5dB ($\alpha = 0.1$, RT60 = 0.2s). In contrast to Figure 4b, the maximum MSD peak has shifted from $\alpha = 0.4$ to $\alpha = 0.5$, although a sizeable secondary peak of 20.9dB is exhibited at $\alpha = 0.4$.

The TDE RMSE for real recordings in an approx. RT60 = 300ms environment is illustrated in Figure 7a. Compared to Figure 5a, the omnidirectional microphone pattern has a similar error rate of 6.8 samples, and this again significantly outperforms the directional patterns, which is consistent with the theoretical results in Figures 4a and 6a.

Figure 7b displays the ICLD MSD for the offset line array, where the trends shown differ in low frequency subbands (<1.5kHz) to results seen in Figure 5b. Most notably, the cardioid directivity pattern exhibits higher MSD across the spectrum, with up to 4dB improvement over the omnidirectional and figure-8 directivity patterns. This suggests that, in real acoustic environments, perhaps an offset line array such as in Figure 3b is better suited to level-based spatial cues than a circular microphone array configuration.

5. CONCLUSION

The use of speaker location information ('spatial cues') has been shown to effectively segment meeting audio recordings for improved user access and higher level processing and analysis. Meetings have traditionally deployed omnidirectional microphone arrays, and the aim of this paper is to determine what microphone directivity patterns are best suited to different types of spatial cues. Circular and offset line microphone array configurations were evaluated to determine the effect of directivity pattern and varied microphone array configuration; experiments were conducted on theoretical simulations in addition to recordings made in a real reverberant room.

A commonly utilised localisation cue, Time-Delay Estimation (TDE), was found to perform best with an omnidirectional directivity pattern, independent of the varied microphone array configurations. To complement the TDE, an inter-microphone level-based cue was also studied, which responded best to unidirectional (cardioid-type) patterns in low reverberation and increasingly bidirectional (figure-8) patterns in high reverberation. Additionally, the cardioid directivity pattern performed better with the offset line array over the circular microphone configuration with the real recordings, despite exhibiting similar results in the theoretical simulations. This result suggests that an offset line array configuration may be better suited to spatial cue extraction in real multiparty meeting environments. More extensive experiments are currently underway to investigate these findings and investigate other array configurations.

REFERENCES

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

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

[3] G. Lathoud, I. McCowan, D. Moore, "Segmenting Multiple Concurrent Speakers using Microphone Arrays," in proc. *Eurospeech 2003*, pp. 2889-2892, Geneva, Sept. 2003.

[4] Y. Tamai, S. Kagami, H. Mizoguchi, K. Sakaya, K. Nagashima, T. Takano, "Circular Microphone Array for Meeting System," in proc. *IEEE Sensors*, vol. 2, pp. 1100-1105, Oct. 2003.

[5] X. Chen, Y. Shi, W. Jiang, "Speaker Tracing and Identifying based on Indoor Localization System and Microphone Array," in proc. *IEEE AINAW '07*, pp. 347-352, May 2007.

[6] E. Cheng, I. Burnett, C. Ritz, "Varying Microphone Patterns for Meeting Speech Segmentation using Spatial Audio Cues," in proc. *Pacific-Rim Conference on Multimedia (PCM '06)*, Hangzhou, China, Nov. 2006.

[7] J. DiBiase, H. Silverman, M. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward (eds.), pp. 157-180, Berlin: Springer-Verlag, 2001.

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

[9] J. Breebaart et al., "High Quality Parametric Spatial Audio Coding at Low Bitrates," *116th AES Convention*, Berlin, May 2004.

[10] J. Breebaart, C. Faller, Spatial Audio Processing: MPEG Surround and Other Applications, England: John Wiley & Sons Ltd., 2007.

[11] J. Eargle, *The Microphone Book*, 2nd Ed., Burlington, Mass.: Focal Press, 2004.

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