USING SPATIAL CUES FOR MEETING SPEECH SEGMENTATION

E. Cheng, J. Lukasiak, I. S. Burnett, D. Stirling

School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, Australia
[ecc04, jasonl, ianb, stirling]@uow.edu.au

ABSTRACT

This work investigates the validity and accuracy of using spatial cues with Time-Delay Estimation (TDE) as a method of segmenting multichannel recorded speech by speaker location. In environments such as meetings where speakers do not significantly alter position, segmentation by speaker location essentially leads to segmentation by speaker 'turn'. The proposed system calculates location information using TDEs and spatial cues extracted from multichannel meeting audio recordings. This location information is then input into a simple segmentation algorithm. Experiments have been performed on both theoretical and real meeting recordings with non-overlapping speakers, and theoretical recordings with overlapping speakers. Segmentation results reveal the most robust cue to be a combination of spatial information and TDEs. This cue combination leads to greater segmentation accuracy for classifying individual speakers and detecting overlapping sections than using spatial cues or time-delay information alone.

1. INTRODUCTION

Current methods of browsing meeting audio recordings are inefficient and cumbersome; they largely rely on users identifying important sections based on the structure of the recorded signals. Fundamental to making such audio recordings easy to browse, are techniques that automatically segment, annotate, and index recorded speech in a semantically meaningful manner.

This paper focuses on the segmentation of meeting speech based on speaker location. In meeting environments, speakers are generally spatially stationary, and hence location information can be used to segment meeting speech by each speaker's period of participation or 'turn'. This basic information may then be used as a basis for speech annotations such as speaker location, change in speaker or number of speakers (using overlap detection).

Previous work in the specific area of meeting speech segmentation by speaker location was reported by Lathoud et. al. [1, 2] and Ajmera et. al. [3]. These techniques estimate speaker location using Time-Delay Estimates (TDEs) based upon Generalized Cross-Correlation (GCC) between the microphone signals [1, 2], and beamforming methods such as SRP-PHAT [2, 3]. These location estimates are then used as input features to intelligent segmentation models such as Hidden Markov Models [1] and smart clustering [3].

In this paper, we propose an augmented technique that builds upon work in [1, 2]. In particular, additional location information is fed to subsequent segmentation models by complementing the TDE location estimates with more sophisticated spatial audio cues. To evaluate the validity of the spatial/TDE cue combination, a simple, well understood data mining algorithm is used for segmentation.

The advantages of the system proposed in this paper are that both time delay and amplitude levels are used as cues. This approach follows that of [3] which extracts the Mel Frequency Cepstral Coefficients from lapel microphones. In contrast, the level information in this paper is directly extracted from the same microphone array signals used for the TDEs. Psychoacoustic effects are also incorporated into the spatial cue extraction allowing perceptual localization to be exploited in subsequent cue processing.

The body of this paper has a description of the proposed approach in Section 2, followed by implementation details in Section 3. Section 4 details the experiments performed, summarizing the results obtained in Section 5. The paper is then concluded in Section 6.

2. APPROACH

2.1. Localization Cue Extraction

2.1.1. Spatial Cue Estimation

In sound source localization, the key cues used by humans are the inter-aural level and time differences [4]. This concept has been previously used in a low-rate spatial audio coding scheme known as Binaural Cue Coding (BCC) [5, 6]. In BCC, the perceptual, spatial image of multichannel audio is captured by extracting Inter-Channel Level and Time Difference cues (ICLD and ICTD) during analysis [5, 6].

The system proposed in this paper extracts inter-channel spatial cues based on the BCC analysis process [5, 6]. For a C channel input, each input channel c is split into M frames using 50% overlapped Hanning windows. The spectra $X_{c,m}[k]$ for each frame m is then calculated using an FFT operation. BCC then decomposes $X_c[k]$ (where m is omitted for simplicity) into B frequency subbands with bandwidths matching the critical bands of human hearing [5]. The DFT coefficients in each subband b are denoted by $k \in \{A_{b-1}, A_{b-1}+1, \cdots, A_b-1\}$, where A_b are the subband boundaries with $A_0 = 0$.

For each channel pair, p, the cues are then calculated for each subband, b. Mathematically, the ICLD cues are extracted according to [5]:

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

E. Cheng is jointly supported by the CSIRO ICT Centre

Since the localization mechanism in the human hearing system is dependent on frequency [4], the ICTD calculation estimates the average phase delay per subband between channels for frequencies below 1.5kHz, and the group delay between channels at higher frequencies [5]. These ICTD cues were initially extracted but did not contribute positively to the overall set of cues as discussed in Section 5. Thus, a second approach which weights the FFT bins according to magnitude was investigated. This Inter-channel Phase Difference (IPD) cue is obtained in subbands up to 2kHz according to [7]:

$$IPD_p[b] = \angle \left(\sum_{k=A_{b-1}}^{A_b-1} X_1[k] X_2^*[k] \right) \quad (2)$$

2.1.2. Time Delay Estimation

To extract the time delay estimations, the same approach as [1, 2] is used. That is, the Generalized Cross-Correlation with PHAse Transform (GCC-PHAT) is calculated between each channel pair, p. The PHAT weighted GCC is given by [2, 8]:

$$\hat{G}_{X_1 X_2, p}[k] = \frac{X_1[k] \cdot X_2^*[k]}{|X_1[k] \cdot X_2^*[k]|} \quad (3)$$

Using an Inverse Fast Fourier Transform (IFFT), the phase correlation function is derived by:

$$\hat{R}_{12,p}[\tau] = IFFT(\hat{G}_{X_1 X_2, p}) \quad (4)$$

The TDE $\hat{\tau}_{12,p}$ is then extracted by locating the maximum of $\hat{R}_{12,p}$, such that: $\hat{\tau}_{12,p} = \arg\max_{\tau} \hat{R}_{12,p}[\tau]$ (5)

To minimize erroneous TDE values, the search range of delays is constrained to an interval such that [8]: $-D \leq \hat{\tau}_{12,p} \leq D$ (6)

2.2. Segmentation Algorithm

To separate the performance of the proposed cue extraction process from the performance of a complex segmentation model, a simple, single frame-based segmentation algorithm was employed. This ensured that the validity of the cues themselves would be accurately determined.

A decision-tree based approach was adopted for segmentation, the advantage of this approach was that it removed the need for memory within the algorithm, as would be required in Hidden Markov Model approaches (used in [1]). Allowing for a frame-by-frame classification of the cues, the decision tree was a learnt model induced from supervised training data where the total number of speakers (classes) was known. The learning algorithm employed a divide-and-conquer strategy [9] that selected classes together with an appropriate test to best partition the initial mixture of classes into a number of purer subsets. This was achieved through a number of metrics using various forms of relative entropy such as split information and gain ratio [9].

3. MEETING RECORDINGS

To test the cues in an 'ideal' meeting environment, a set of 'theoretical' meeting recordings were generated and processed by the proposed system. For non-overlapping speakers, a second set of simulations using real meeting data were performed to evaluate the robustness of the cues. To enable valid comparisons with previous work, simulations used the same meeting recordings as in Lathoud et. al. [1, 2]. A 6 minute subset of the available corpora

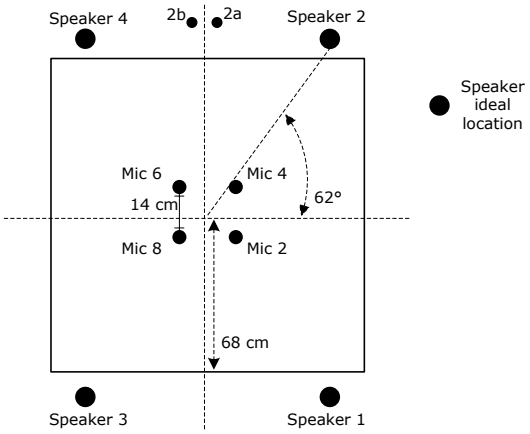

Fig. 1. Meeting Room Configuration, based on [1]

was selected, consisting of 30 speaker turns varying in duration from 5 to 20 seconds with each speaker equally represented.

Both the theoretical and real meeting recordings were sampled at 16kHz. To extract the spatial cues and TDEs, overlapped, windowed frames of length 32ms were processed every 16ms. With the microphone array configuration illustrated in Fig. 1, the maximum delay was less than 1ms. Hence, D in Equation 6 was set to 1.875ms.

3.1. Ideal Meeting Recordings

To generate the 'theoretical' or 'ideal' recordings, the same meeting room configuration used in the real meeting recordings of [1, 2] was adopted (see Fig. 1). Four 'clean' speakers were emulated by taking files from the Australian National Database of Spoken Languages. Each speaker turn was made to be approximately one minute, resulting in a total of about 3.5 minutes of 'meeting'. These four speech files were then 'spatialized' by simulating the meeting room configuration illustrated in Fig. 1. The microphone array signals were synthesized by attenuating the signal amplitudes proportionally to the inverse squared distance between the speaker and microphones, and introducing theoretical time delays relative to that distance. The result is a 'ideal' meeting recording with no reverberation or other 'room' effects.

To generate theoretical meeting recordings with overlapped speakers, dual-speaker overlap was simulated by superposing two speakers using the synthesized array signals. With four speakers in the room, all six dual-speaker combinations were equally represented with single speaker segments of 6 seconds interspersed with overlapped sections of 3 seconds, forming a total 'meeting' of about 3 minutes.

4. EXPERIMENTS

To evaluate the accuracy and robustness of the location cues, a set of simulations were performed. For non-overlapping speakers, the following simulations were conducted:

1. Ideal meeting recordings, speakers at ideal locations (as illustrated in Fig. 1);
2. Real meeting recordings, speakers at ideal locations;
3. Ideal meeting recordings, speakers at non-ideal locations.

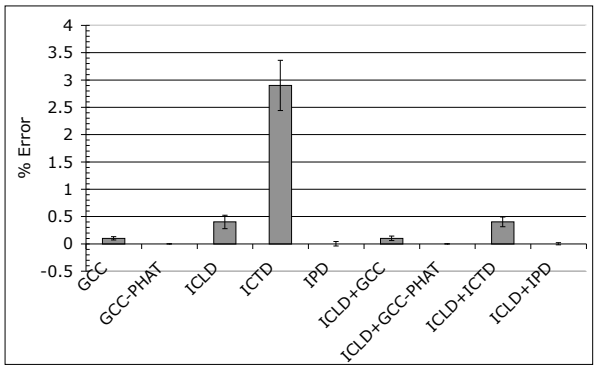

Fig. 2. Theoretical recordings - Non-overlapped speakers

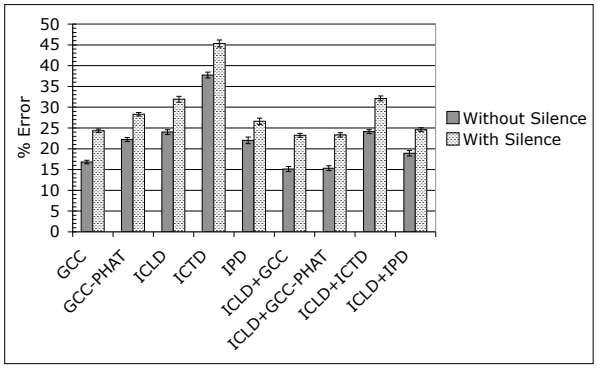

Fig. 3. Real recordings - Non-overlapped speakers

Simulation 3 investigated the effect of spatial variations through using non-ideal speaker locations in an ideal environment. Training data consisted of each speaker located in their quadrant (see Fig. 1), at 10 positions equally spaced $\pi/20$ radians apart on a circle of radius 68cm around the microphone array. Training the decision tree with these 40 locations, test data was then generated with each speaker placed at a fixed location.

To evaluate the system with overlapping speakers, Simulation 4 tested the robustness of all the cues against theoretical meeting recordings with overlapped dual-speaker segments.

The evaluation criteria was the error rate, defined as the percentage of incorrectly segmented frames over total number of frames. In simulations where the training and test data were cues calculated from the same audio files, the data was split using a 10-fold cross validation process; a standard approach taken in data mining [9]. The data was divided into 90% training and 10% testing, with each speaker or overlap scenario sampled in proportion to its occurrence. This data division was repeated 10 times (folds), each time on a different training/test set. An average error rate was returned and a 95% confidence interval calculated. For simulations where the training and test data contained cues from different audio files, the error rate returned from the decision tree was used.

5. RESULTS AND DISCUSSION

5.1. Speakers at Ideal Positions

Results from Simulation 1 are shown in Fig. 2, suggesting that all the location cues except the ICTD are valid for speech segmentation by location. In ideal conditions, the poor performance of the ICTD compared to all the other cues indicates that the ICTD may not be suitable for speech segmentation. In the ICTD estimation, performed as defined in BCC [5, 6], the phase differences between channels are averaged in each subband. However, the addition of phase is invalid as it combines terms from different frequencies. Furthermore, the BCC technique employed applies no weighting to the FFT bins. Hence, bins with very low magnitudes (and with phase uniformly distributed between $\pm\pi$ [10]) corrupt the ICTD estimation. As seen in Fig. 2, the weighting used in IPD calculations reduces this problem, making the IPD a more reliable spatial cue for segmentation.

The set of results from Simulation 2 are shown in Fig. 3, using real meeting recordings with and without silence segments. Removing silence reduces the location ambiguity in the signals since no speakers are active during these periods. The poor performance of the ICTD compared to all the other cues is again evident in this simulation. However, by combining various time delay cues (GCC, GCC-PHAT, ICTD) and IPD with ICLD, segmentation results improve dramatically over using the cues alone. ICLD combined with IPD is slightly poorer than ICLD with GCC and GCC-PHAT, showing that subband phase estimates are not as accurate as TDE. These results indicate that ICLDs used in conjunction with TDEs form good cues for location segmentation, and appear robust against the effects of a real meeting environment.

In Fig. 3, the GCC-PHAT does not perform as well as GCC for the tested data set. This contradicts past work which states that GCC-PHAT performs better in environments with reverberation [8]. Analysis of the GCC and GCC-PHAT illustrated that the GCC-PHAT does give a more impulsive time-domain waveform as expected. However, GCC-PHAT is generally performed using longer signal frames than we use here, as short-time frames degrade performance as the signal statistics fluctuate [8].

5.2. Speakers at Non-Ideal Positions

The set of results from Simulation 3 is shown in Table 1, where three test speaker positions were used:

3.1 Speaker at ideal location;

3.2 Speaker at own quadrant's edge (e.g. position 2a for Speaker 2 in Fig. 1) - worst case scenario for location ambiguity between two quadrants;

3.3 Speaker at neighbour's quadrant edge (e.g. position 2b for Speaker 2 in Fig. 1).

Clearly, results in the 'Ideal' column of Table 1 show that all cues perform well for ideal speaker locations. In particular, the ICLD and ICTD improve over the corresponding results in Fig. 2, while GCC, GCC-PHAT and IPD are unaffected. This suggests that providing the valid set of cue values for each speaker enhances the segmentation algorithm accuracy.

Results for the worst-case speaker location are shown in the 'Edge' column of Table 1. Again, GCC, GCC-PHAT and IPD are unaffected by this worst-case scenario which may give ambiguous location cues. Only the ICLD and ICTD slightly deteriorate, indicating that all cue combinations are robust against non-ideal speaker locations.

Results in the 'Next Quadrant' column of Table 1 indicate that all cues correctly classify the speaker location, even though the

Table 1: Theoretical recordings - Non-ideal speaker location

% Error	Ideal	Edge	Next Quadrant
GCC	0.1	0.1	99.9
GCC-PHAT	0	0	100
ICLD	0.2	1.8	99
ICTD	1.2	2.6	97.9
IPD	0	0	100
ICLD + GCC	0.1	0.1	99.9
ICLD + GCC-PHAT	0	0	100
ICLD + ICTD	0.2	1.8	99
ICLD + IPD	0	0	100

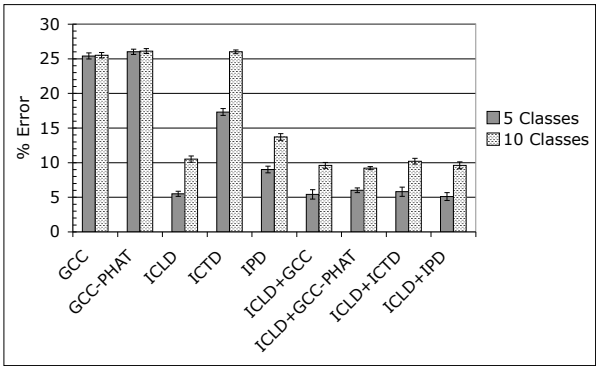

Fig. 4. Theoretical recordings - Overlapped speakers

speaker does not belong to that quadrant. This suggests that location information, rather than speaker dependent characteristics, dominates the segmentation process. Again, the ICTD and ICLD perform only slightly worse than all the other cues, while the IPD, GCC and GCC-PHAT continue trends as the most robust cues in this simulation.

5.3. Overlapped Speakers

Fig. 4 shows the results for Simulation 4. One experiment used 5 decision tree classes; one class for each speaker, and one for the overlapped segments. The second experiment used 10 decision tree classes; one for each speaker, and one for each dual-speaker combination. Using 10 classes produced poorer segmentation results than using 5 classes, potentially due to less training data available as the same overlapped data is now split into 6 separate classes. The GCC and GCC-PHAT perform very poorly compared to the spatial cues, because the cross-correlation method chooses the strongest speaker at each time-frame. Attempting to map a dual-speaker segment onto a single peak-based correlation measure, as in GCC and GCC-PHAT, will result in high error rates. The spatial cues (ICLD, ICTD, IPD) perform significantly better than the TDEs as they exploit subband cue analysis and hence spectral diversity, to provide enhanced detection of multiple speakers. Consistent with single-speaker simulations, the combination of ICLD with TDEs or IPD provides the lowest segmentation error.

6. CONCLUSION

This paper proposed a new approach to applying location information from multichannel meeting speech recordings for segmentation purposes. In addition to extracting time delay estimations using traditional cross-correlation techniques, this paper complemented these with spatial cue estimations. Simulations on an ideal meeting environment showed that TDEs are well matched to location-based segmentation. However, in real meeting environments, combining the TDEs with spatial level differences significantly improved the segmentation results. This cue combination proved robust against non-ideal speaker locations, speaker-dependent characteristics and performed significantly better than TDE techniques alone in detecting overlapped speakers. In particular, combining the spatial level differences with the GCC TDE proved most robust across these practical meeting conditions. Furthermore, results illustrated that subband phase techniques without frequency bin weighting are not suitable for the purposes of segmentation by location.

This paper has shown that combining spatial level differences with GCC gives the best location information for segmentation. For future work, improved accuracy can be obtained by employing more sophisticated segmentation algorithms including using memory to remove rapid speaker transitions.

7. REFERENCES

[1] G. Lathoud and I. A. McCowan, "Location based speaker segmentation," in *ICASSP '03*, vol. 1, pp. 176–179, Hong Kong, 2003.

[2] G. Lathoud, I. A. McCowan, and D. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *Eurospeech '03*, 2003.

[3] J. Ajmera, G. Lathoud, and I. A. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *ICASSP '04*, vol. 1, pp. 605–608, Montreal, 2004.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, 1997.

[5] C. Faller and F. Baumgarte, "Binaural cue coding-part ii: Schemes and applications," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 520–531, 2003.

[6] C. Faller and F. Baumgarte, "Binaural cue coding applied to stereo and multi-channel audio compression," in *AES 112th Conv. Paper 5574*, Munich, 2002.

[7] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "High-quality parametric spatial audio coding at low bitrates," in *AES 116th Conv. Paper 6072*, Berlin, 2004.

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

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, San Francisco, 2000.

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.