LOCATION BASED SPEAKER SEGMENTATION

Guillaume Lathoud *Iain A. McCowan*

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P. O. Box 592, CH-1920 Martigny, Switzerland
{lathoud, mccowan}@idiap.ch

ABSTRACT

This paper proposes a technique that segments audio according to speakers based on their location. In many multi-party conversations, such as meetings, the location of participants is restricted to a small number of regions, such as seats around a table, or at a whiteboard. In such cases, segmentation according to these discrete regions would be a reliable means of determining speaker turns. We propose a system that uses microphone pair time delays as features to represent speaker locations. These features are integrated in a GMM/HMM framework to determine an optimal segmentation of the audio according to location. The HMM framework also allows extensions to recognise more complex structure, such as the presence of two simultaneous speakers. Experiments testing the system on real recordings from a meeting room show that the proposed location features can provide greater discrimination than standard cepstral features, and also demonstrate the success of an extension to handle dual-speaker overlap.

1. INTRODUCTION

Segmentation of audio into speaker turns is an important task in many speech processing applications. It can be a necessary preprocessing task for speech and speaker recognition systems, and also a means of identifying higher level structure in audio documents, such as periods corresponding to monologues, dialogues, or general discussion.

One of the issues in designing a speaker segmentation system is the choice of features capable of distinguishing individual speakers. Mel-frequency or linear predictive cepstral coefficients are most commonly used, either in an acoustic change detection or speaker clustering framework [1, 2, 3]. The performance of these features can be limited in practice, as they tend to discriminate only between speaker classes, rather than individual speakers.

In this paper, we investigate the use of location-based features to distinguish between speakers for segmentation. In many multi-party conversations, such as meetings and teleconferences, the location of a speaker remains stationary throughout most of the conversation. In situations where it is practical to acquire the audio across multiple microphones, microphone array processing techniques may be used to dynamically estimate the location of dominant speech sources [4]. In such cases, the location estimates should discriminate between speakers, and hence could be used as features in a segmentation framework.

In the present work, we estimate time delays from the generalised cross-correlation between paired microphones within an array. These estimated time delays form input features that are integrated in a GMM/HMM framework to segment the audio according to a set of discrete speaker locations. The discrimination provided by the location features, coupled with the HMM's ability to model sequences, makes it possible to extend the system to segment the conversation in terms of higher level structure. To demonstrate this, we propose an extension to handle the case of overlapping speech from multiple simultaneous speakers. Such speaker overlap has been identified as a significant problem for speech segmentation and recognition of multi-party conversations [5].

The proposed location-based speaker segmentation system is assessed on real recordings from a 4-element microphone array in a meeting room. Results are presented comparing the performance of the location features to standard cepstral (LPCC) features for single speaker segments. In addition, experiments on overlapping speech segments demonstrate the success of the proposed extension to handle dual-speaker overlap.

2. PROPOSED SYSTEM

In this section we detail the proposed system for location-based speaker segmentation. A feature space is defined as the set of time delay estimates (TDE's) across M pairs of microphones. Each time delay estimate measures the difference in the time of arrival between the signals on a microphone pair. Gaussian distributions are used to model the behaviour of the observed features around a number of speaker locations. These then form the state distributions in an HMM, which can be used to obtain a maximum likelihood segmentation into speaker turns.

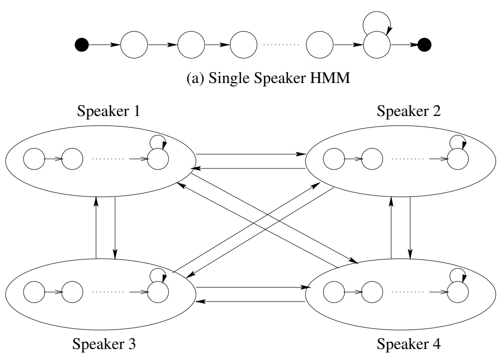
2.1. Model Formulation

As the basis of our model, we assume that a speaker k is confined to a physical region centred at location $\mathbf{x}_k \epsilon \Re^3$.

We define the vector of theoretical time delays $\boldsymbol{\mu}_k$ associated with the speaker location $\mathbf{x}_k$ as :

$$\begin{aligned} \boldsymbol{\mu}_k &= f(\mathbf{x}_k) \\ &= \begin{bmatrix} \mu_k^{(1)} & \mu_k^{(2)} & \cdots & \mu_k^{(M)} \end{bmatrix}^T \end{aligned}$$

This work was carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2, and the European project M4 through the Swiss Federal Office for Education and Science.



(b) HMM Topology for Segmentation System

Fig. 1. HMM topology

where $\mu_k^{(m)}$ is the time delay (in samples) between the microphones in pair m, given by

$$\mu_k^{(m)} = \frac{\left(||\mathbf{x}_k - \mathbf{m}_1^{(m)}|| - ||\mathbf{x}_k - \mathbf{m}_2^{(m)}||\right) f_s}{c}$$

where $\mathbf{m}_1^{(m)}$ and $\mathbf{m}_2^{(m)}$ are the locations of the microphones in pair m, $||\cdot||$ is the Euclidean norm, and f_s is the sampling frequency.

Given an input observation vector $\hat{\mathbf{D}}_t$ of time delay estimates $\hat{\tau}_t^{(m)}$ at time t, we model the distribution of the observation given the speaker at location $\mathbf{x}_k$ as :

$$p(\hat{\mathbf{D}}_t|\mathbf{x}_k) = \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\Sigma}_k$ is the covariance matrix. The Gaussian distribution is used to model the effects of variations in speaker location around $\mathbf{x}_k$, as well as uncertainty in the observed time delay estimates due to reverberation and noise.

In this paper, each time delay estimate $\hat{\tau}_t^{(m)}$ is calculated from the generalised cross-correlation (GCC) [6] between the signals captured on the m^{th} microphone pair. A phase transform (PHAT) weighting function is applied in order to improve the robustness of the estimates to reverberation. In addition, time-domain interpolation of the GCC function is performed in order to achieve sub-sample precision. Full details of the GCC-PHAT time delay estimation procedure can be found in [7].

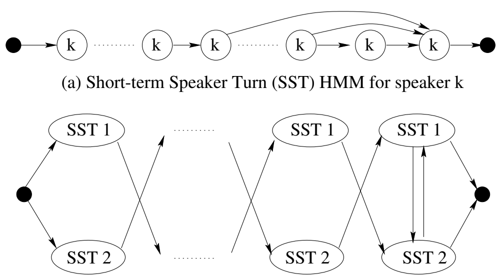
2.2. HMM Segmentation Framework

To segment the audio signal according to speaker turns, we use a HMM framework similar to that proposed in [8] for speech/music segmentation.

We define a minimum duration left-to-right HMM for each speaker k, where all states are modeled using the Gaussian density $p(\hat{\mathbf{D}}_t|\mathbf{x}_k)$ as proposed in the previous section. This single speaker HMM topology is shown in Figure 1(a).

A grammar is introduced to define transitions between speakers, excluding self-loops. The resulting HMM for the segmentation system is shown in Figure 1(b) for the case of $K = 4$ speakers.

Given an observation sequence of feature vectors $\hat{\mathbf{D}}_{1:t}$, the optimal path through the HMM can be found using Viterbi decoding, giving the maximum likelihood segmentation in terms of speaker locations.



(b) Dual-speaker HMM (speakers 1 and 2)

Fig. 2. Dual-speaker HMM

3. EXTENSION TO SEGMENTS OF SPEAKER OVERLAP

In this section, we propose an extension to handle segments containing two overlapping speakers. We first discuss the occurrence of overlapping speech in multi-party conversations, and the problem this poses for source localisation and speaker segmentation systems. We then propose a dual-speaker HMM topology which can be used to extend the system of Section 2.2 to handle segments of overlapping speech.

3.1. Overview of Problem

Overlapping speech is a common problem in multi-party conversations, such as meetings and telephone conversations. Overlap may occur when someone attempts to take over the main discussion, when someone interjects a brief comment over the main speaker, or when a separate conversation is taking place in addition to the main discussion. In [5] it was identified that around 10-15% of words, or 50% of contiguous speech segments, in a meeting or telephone conversation contain some degree of overlapping speech.

These overlap segments are problematic for speech recognition, producing an absolute increase in word error rate of between 15-30% using close-talking microphones for a large vocabulary task [5, 9]. For applications that involve meetings or teleconferences, it is thus important to not only segment the audio into single speaker turns, but also to identify segments of overlapping speech along with their constituent speakers. This is a difficult problem for standard speaker segmentation and speech activity techniques using close-talking microphones [10, 9].

3.2. Proposed Dual-Speaker HMM

If there are K individual speakers, an overlap segment may be defined as one in which there are n active speakers, where $2 \leq n \leq K$. We restrict this current work to the case of $n = 2$ which we will refer to as dual-speaker overlap.

Empirical observation of TDE values during segments of overlapping speech shows an alternating sequence of short-term speaker turns (SST's). These SST's are due to frame-by-frame variations in relative energy levels between the two speakers, as the TDE features are computed from the highest energy GCC peak in each frame. To model this behaviour, we first define a left-to-right HMM that represents a SST, shown in Figure 2(a). This model imposes a minimum duration to exclude noise, as well as a maximum duration to exclude single-speaker segments.

From this, a dual-speaker HMM is then proposed as an alternating sequence between the SST models of two speakers, as shown in Figure 2(b). Similar to the single speaker HMM, a minimum duration constraint is included to eliminate undesired short segments.

Subsequently, an audio signal containing a series of single speaker and dual-speaker segments may be segmented using :

- K single speaker HMMs, as shown in Figure 1(a), and
- $K(K-1)/2$ dual-speaker HMMs, as shown in Figure 2(b).

These single and dual-speaker classes are combined in an inter-class grammar that forbids self-loops, similar to that shown in Figure 1(b) for the single speaker case.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Configuration

Experiments were conducted in a meeting room using a 4-element microphone array ($M = 6$ pairs) placed in the centre of a table, with speakers seated at 4 different locations around the table, as shown in Figure 3. A test database was recorded simultaneously across all microphones at a sampling rate of 16 kHz. The total database duration was 20 minutes, consisting of 5 minutes of speech from a different person for each location.

These four 5 minute single speaker/location files were randomly recombined to form two separate test sets. *Test set 1 (non-overlap)* contained only single speaker segments without any overlap segments. Nine files containing 10 speaker turns were constructed in a random manner, with segments varying from between 5 to 20 seconds in duration. *Test set 2 (overlap)* was constructed from the same database in a similar manner, however this time a short overlap segment was included at each speaker change. The test set consisted of six files, each containing 10 single speaker segments (of between 5-17 seconds duration), interleaved with 9 segments of dual-speaker overlap (of between 1.5-5 seconds duration). The TDE features were calculated on 32ms input frames, every 16ms.

4.2. Evaluation Criteria

To assess the system performance, the following metrics were used

- frame accuracy (FA) :

$$\frac{\text{number of correctly labelled frames}}{\text{total number of frames}} \times 100\%$$

- precision (PRC) :

$$\frac{\text{number of correctly found segment boundaries}}{\text{number of segment boundaries detected}}$$

- recall (RCL) :

$$\frac{\text{number of correctly found segment boundaries}}{\text{number of true segment boundaries}}$$

The precision and recall values were combined in a single metric using the common F-measure [3], which is defined as :

$$F = 2 \times \text{PRC} \times \text{RCL}/(\text{PRC} + \text{RCL})$$

and varies between 0 and 1. In most cases, a short interval of silence exists between two consecutive speech segments, and so in comparing segment boundaries to the ground truth, a tolerance interval of ± 1 second was chosen.

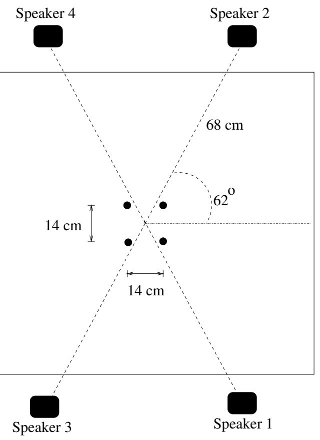


Fig. 3. Experimental setup

system	FA	PRC	RCL	F
TDE features	99.1%	0.98	0.98	0.98
LPCC features	88.3%	0.81	0.73	0.77

Table 1. Results for single speaker system (test set 1)

4.3. Results and Discussion

4.3.1. Single Speaker System

A first series of tests was conducted to investigate the performance of the proposed location-based segmentation on the non-overlap test set, comparing it to an equivalent (single-channel) system using standard linear predictive cepstral coefficients (LPCC's).

The HMM consisted of 4 single speaker classes, as shown in Figure 1(b). The distribution means were set as defined in Section 2.1, and all variances were set to the same value (unity), as speaker regions were of uniform size. A minimum duration of 2 seconds was imposed for each speaker class, with the final class state having a self-loop probability of 0.9. Transitions in the inter-class grammar were equally weighted.

To provide a basis for comparison, an equivalent system was implemented using standard LPCC's of dimension 12. For the location-based system, knowledge of the speaker regions represents *a priori* information of the state distributions. For this reason, individual speaker GMM's (using 8 mixtures) were trained for the LPCC system using separate training data for each speaker. These trained GMM's then formed the state distributions in the same HMM topology as was used for the location-based system.

Table 1 presents the results for the location- and LPCC-based single speaker systems. These results show the improved discrimination provided by the location-based features, as well as the suitability of the proposed HMM framework for segmentation. The improved results of the location-based system are achieved with lower model complexity (one mixture per GMM, compared to 8 for LPCC's), as well as simpler training, through direct calculation of the state distribution associated with each location.

test set	FA	PRC	RCL	F
non-overlap (1)	99.1%	0.98	0.98	0.98
overlap (2)	94.1% (85.5%)	0.94	0.86	0.90

Table 2. Results for extended system (test sets 1 and 2). The FA calculated only on actual overlap segments is shown in parentheses

4.3.2. Extended System including Dual-Speaker Overlap

A second series of tests was conducted to evaluate the performance of the proposed extension to include dual-speaker overlap segments.

For this scenario, the HMM topology from the previous experiments was extended by adding the 6 dual-speaker classes, as defined in Section 3. Each short-term speaker turn (SST) was constrained to a duration of 3-10 frames. These SST's were then combined in a minimum duration sequence of 1 second. Once again, transitions in the inter-class grammar were all equally weighted.

As this topology was designed directly from observations of the temporal behaviour of the time delay features during overlap segments, direct comparison with the LPCC features was considered inappropriate in this case. This extended system was tested on both non-overlap and overlap test sets, with results shown in Table 2.

We first observe that the results on test set 1 using the extended system are identical to those obtained using only the 4 single speaker classes, indicating that the addition of the 6 dual-speaker overlap classes does not affect the system's ability to discriminate single speaker segments.

Secondly, we see that a high frame accuracy and F-measure are obtained on the overlap test set. This indicates both the suitability of the proposed overlap class topology, as well as the power of the HMM to represent more complex segment structure. We note that part of the decrease in FA for overlap segments may be attributed to the shorter segment duration and difficulty in defining a precise ground truth. While we have only investigated the case of dual-speaker overlap in our experiments, the HMM system has the potential to segment a multi-party conversation in terms of higher level structure, for example according to presentations, dialogues or general discussion (see e.g. [11]). We also note that, while the state distributions and HMM topologies used in these experiments have been explicitly designed, this information could also be learned directly from data in a supervised or unsupervised manner.

5. CONCLUSIONS

This paper has proposed a framework for speaker segmentation based on location information. The speaker location is modelled using time delays between microphone pairs from a microphone array. These time delays form the input feature vectors to a GMM-HMM system. The optimal path found by decoding an input sequence results in a maximum likelihood segmentation of the audio according to the speaker locations. Experiments on single speaker segments show that the proposed location-based features provide greater discrimination than standard cepstral parameters.

In addition, an extension to handle the case of overlapping speech segments is proposed. A dual-speaker overlap HMM topology is investigated, and is shown to provide high segmentation accuracy in a second set of experiments. These results are significant, as they show that the system is not only capable of distinguishing regions of overlapping speech, but also identifying the constituent speakers.

An obvious limitation of the system is that it assumes each speaker is associated with one location (and vice-versa). This could be addressed by combining this technique with traditional cepstral based speaker clustering or recognition systems. Other continuing work seeks to test the system on real multi-party conversational speech recorded in meetings, and investigate extensions such as unsupervised clustering of speaker locations and online adaptation of speaker priors.

6. ACKNOWLEDGEMENTS

The authors wish to thank Daniel Gatica-Perez and Darren Moore for their advice and comments on the work presented in this article.

7. REFERENCES

[1] S. Chen and P. Gopalkrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. *IBM Technical Journal*, 1998.

[2] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan. Unknown-multiple speaker clustering using HMM. In *Proceedings of ICSLP-2002*, 2002.

[3] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *Proceedings of ICASSP-2000*, 2000.

[4] J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 8, pages 157–180. Springer, 2001.

[5] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proceedings of Eurospeech 2001*, volume 2, pages 1359–1362, 2001.

[6] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustic, Speech and Signal Processing*, ASSP-24(4):320–327, August 1976.

[7] M. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of ICASSP-96*, 1996.

[8] J. Ajmera, I. McCowan, and H. Bourlard. Robust HMM-based speech/music segmentation. In *Proceedings of ICASSP-02*, 2002.

[9] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Proceedings of Human Language Technology Conference*, 2001.

[10] T. Pfau, D. Ellis, and A. Stolcke. Multispeaker speech activity detection for the ICSI meeting recorder. In *Proceedings of ASRU-01*, 2001.

[11] D. Zotkin, R. Duraiswami, and L. Davis. Multimodal 3-d tracking and event detection via the particle filter. In *Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, 2001.