Segmenting Multiple Concurrent Speakers Using Microphone Arrays

Guillaume Lathoud, Iain A. McCowan, Darren C. Moore

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

Abstract

Speaker turn detection is an important task for many speech processing applications. However, accurate segmentation can be hard to achieve if there are multiple concurrent speakers (overlap), as is typically the case in multi-party conversations. In such cases, the location of the speaker, as measured using a microphone array, may provide greater discrimination than traditional spectral features. This was verified in previous work which obtained a global segmentation in terms of single speaker classes, as well as possible overlap combinations. However, such a global strategy suffers from an explosion of the number of overlap classes, as each possible combination of concurrent speakers must be modeled explicitly. In this paper, we propose two alternative schemes that produce an individual segmentation decision for each speaker, implicitly handling all overlapping speaker combinations. The proposed approaches also allow straightforward online implementations. Experiments are presented comparing the segmentation with that obtained using the previous system.

1. Introduction

Segmenting the speech signal in terms of speaker turns is a necessary pre-processing task for many applications: speech recognition needs segments of short length, and browsing of recordings is made easier with a timeline showing who is speaking and when. Other applications include broadcast news indexing, meeting summarisation and video surveillance.

While traditional audio features (LPCC, MFCC, energy, etc.) have been used successfully on broadcast recordings and telephone speech, multi-party conversations such as meetings present a more difficult case due to the high amount of overlapping speech in spontaneous conversations [1]. It is difficult to resolve overlaps when using single microphone techniques, since speech from more than one simultaneous speaker is often recorded by the same microphone (crosstalk phenomenon) [2].

In applications involving multi-party conversations, it may be possible to acquire the speech using microphone arrays. By spatially sampling an acoustic field, microphone arrays provide the ability to discriminate between sounds based on their source location. This directional discrimination can be exploited to enhance a signal from a given location, or simply to locate principal sound sources in the field.

In [3], we introduced an approach that processed location-based features from a microphone array within a GMM/HMM framework to produce a global segmentation of speaker turns. The approach gives accurate segmentation on test data including segments with two simultaneous speakers. However, it suffers from the limitation that each possible combination of active speakers (including overlap) has to be modeled with a separate HMM, leading to $(2^K - 1)$ classes, where K is the number of speakers.

In this work, instead of performing a global segmentation in terms of all possible single and multiple speaker classes, we propose two techniques that produce K parallel individual speaker segmentations. In this way, the need to define all possible combinations of active speakers is removed, and any number of concurrent speakers is handled implicitly.

In experiments, results are compared to those obtained using the previous approach, demonstrating that both new approaches successfully handle both single speaker and dual-speaker overlap cases.

Section 2 introduces the fundamentals of localisation using microphone arrays. Section 3 describes the two proposed approaches that address the limitation of the previous approach. Section 4 presents the experiments and a discussion of the results obtained.

2. Localisation Fundamentals

This section recalls the non-linear relationship between physical space and time-delay space, and then summarises the Generalized Cross-Correlation method for time-delay estimation in the case of the PHAse Transform (GCC-PHAT) [4]. We selected the PHAT because it is efficient in high-SNR, reverberant environments such as meeting rooms.

2.1. Link Between Location and Theoretical Time-Delays

We define the vector of theoretical time delays $\boldsymbol{\mu}_k \in \mathbb{R}^P$ associated with the speaker location $\mathbf{x}_k \in \mathbb{R}^3$ as :

$$\begin{aligned} \boldsymbol{\mu}_k &\triangleq f(\mathbf{x}_k) && (1) \\ &\triangleq \left[\begin{array}{ccccc} \mu_k^{(1)} & \cdots & \mu_k^{(p)} & \cdots & \mu_k^{(P)} \end{array} \right]^T && (2) \end{aligned}$$

where P is the number of microphone pairs and $\mu_k^{(p)}$ is the theoretical time delay (in samples) between the microphones in pair p, given by

$$\mu_k^{(p)} \triangleq \frac{\left(||\mathbf{x}_k - \mathbf{m}_1^{(p)}|| - ||\mathbf{x}_k - \mathbf{m}_2^{(p)}|| \right) f_s}{c} \quad (3)$$

where $\mathbf{m}_1^{(p)}$ and $\mathbf{m}_2^{(p)}$ are the locations of the microphones in pair p, $||\cdot||$ is the Euclidean norm, f_s is the sampling frequency, and c the speed of sound in the air (usually 342 m/s).

2.2. GCC-PHAT time-delay estimation

Full details of this procedure can be found in [5]. From two signals $s_1^{(p)}(t)$ and $s_2^{(p)}(t)$ of a given microphone pair p, GCC-

PHAT is defined as:

$$\hat{G}^{(p)}_{PHAT}(f) \triangleq \frac{S^{(p)}_1(f) \cdot \left[S^{(p)}_2(f)\right]^*}{\left|S^{(p)}_1(f) \cdot \left[S^{(p)}_2(f)\right]^*\right|} \quad (4)$$

where $S^{(p)}_1(f)$ and $S^{(p)}_2(f)$ are Fourier transforms of the two signals and $[\cdot]^*$ denotes the complex conjugate. Typically the two Fourier transforms are estimated on Hamming-windowed segments of 20 to 30 ms.

The time-delay estimate (TDE) for the microphone pair p is then defined as:

$$\hat{\tau}^{(p)} \triangleq \arg\max_{\tau} \left(\hat{R}^{(p)}_{PHAT}(\tau)\right) \quad (5)$$

where $\hat{R}^{(p)}_{PHAT}(\tau)$ is the Inverse Fourier Transform of the GCC-PHAT function $\hat{G}^{(p)}_{PHAT}(f)$.

By applying this process for each microphone pair, we construct a vector of TDEs:

$$\hat{\mathbf{D}} \triangleq \left[\ \hat{\tau}^{(1)} \ \ \cdots \ \ \hat{\tau}^{(p)} \ \ \cdots \ \ \hat{\tau}^{(P)}\ \right]^T \quad (6)$$

3. Proposed Approaches

This section presents two new segmentation approaches based on the speaker location. As discussed above, in contrast to [3], they avoid the need for explicit modeling of each overlap class by providing parallel segmentations for each speaker, implicitly handling all overlap combinations. As such, the computational load becomes linear in the number of speakers, rather than exponential.

As in [3], our model assumes that a speaker k is confined to a physical region centred at location $\mathbf{x}_k \in \mathbb{R}^3$. The two approaches presented in the following subsections unfold in two steps:

1. Classify each (speaker, frame) as speech or silence, *independently of other speakers and other frames*, thus obtaining K binary series

$$ss^{(k)} = (ss^{(k)}_1, \ldots, ss^{(k)}_n, \ldots, ss^{(k)}_N)$$

   where k is the speaker index ($1 \leq k \leq K$), n the frame index ($1 \leq n \leq N$) and $ss^{(k)}_n \in \{0, 1\}$. "0" denotes a silent frame, "1" denotes a speech frame.

2. For each speaker k, apply a simple dilation/erosion process to smooth the binary sequence $ss^{(k)}$. This operation aims at connecting frames belonging to the same utterance, as well as eliminating spurious speech segments less than a specified minimum duration.

While the features and models used in the first step differ between the two approaches, the second step is the same for both. We describe the first step in Section 3.1. The dilation/erosion process common to both approaches is described in Section 3.2.

3.1. Step One: Frame-Level Speech/Silence Classification

3.1.1. Speech/Silence Ratio Approach

The features used here are equivalent to those used in the HMM approach presented in [3], i.e. GCC-PHAT TDEs, as defined in (5). For each frame $n \in [1 \ldots N]$, we extract a vector $\hat{\mathbf{D}}_n$ of TDEs. For a given speaker k and a given frame n, we model the likelihood of the observed TDEs with two possible pdfs:

- Speech:

$$p^{k,n}_{speech}(\hat{\mathbf{D}}_n|\mathbf{x}_k) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{k}}, \boldsymbol{\Sigma}_k) \quad (7)$$

  where $\boldsymbol{\Sigma}_k$ is the covariance matrix (typically diagonal) and n the frame index. The Gaussian distribution models the effects of variations in speaker location around $\mathbf{x}_k$, as well as uncertainty in the observed TDEs due to reverberation and noise.

- Silence:

$$p_{silence}(\hat{\mathbf{D}}_n|\mathbf{x}_k) = \frac{1}{\prod_{p=1}^{P} 2\,\tau^{(p)}_{max}} \quad (8)$$

  where $\tau^{(p)}_{max}$ is the maximum time-delay (in samples) between the microphones in pair p and n the frame index. $\tau^{(p)}_{max}$ is directly proportional to the distance between the two microphones:

$$\tau^{(p)}_{max} \triangleq \frac{||\mathbf{m}^{(p)}_1 - \mathbf{m}^{(p)}_2||\ f_s}{c} \quad (9)$$

We can then define the Speech/Silence Ratio (SSR) as:

$$SSR(k,n) \triangleq \frac{p^{k,n}_{speech}(\hat{\mathbf{D}}_n|\mathbf{x}_k)}{p_{silence}(\hat{\mathbf{D}}_n|\mathbf{x}_k)} \quad (10)$$

For a given speaker k and a given frame n, speech/silence classification then amounts to:

$$ss^{(k)}_n = \begin{cases} 0 & \text{if} \quad SSR(k,n) < 1 \\ 1 & \text{if} \quad SSR(k,n) \geq 1 \end{cases} \quad (11)$$

3.1.2. Steered Response Power Approach

In contrast to the single stream of features used in the HMM and SSR approaches, we use here a separate stream of features for each speaker. Therefore, multiple speakers can be active within the same frame. For a given speaker k and a given frame n, we estimate the Steered Response Power (SRP) using a measure known as SRP-PHAT [6]. We sum the time domain version of the GCC-PHAT function defined in (4) at the theoretical time-delays associated with location $\mathbf{x}_k$:

$$P_{SRP}(k,n) \triangleq \frac{1}{P}\sum_{p=1}^{P} \hat{R}^{(p)}_{PHAT}\left(\mu^{(p)}_k\right) \quad (12)$$

where P is the number of microphone pairs and $\hat{R}^{(p)}_{PHAT}(\tau)$ is the time-domain GCC-PHAT. We have the property $P_{SRP}(k,n) \in [-1,+1]$. The higher the value of $P_{SRP}(k,n)$, the more likely it is for speaker k to be active at frame n.

For a given speaker k and a given frame n, speech/silence classification then amounts to:

$$ss^{(k)}_n = \begin{cases} 0 & \text{if} \quad P_{SRP}(k,n) < \theta_{SRP} \\ 1 & \text{if} \quad P_{SRP}(k,n) \geq \theta_{SRP} \end{cases} \quad (13)$$

where $\theta_{SRP} \in [-1,+1]$ is a threshold value that has to be tuned. In practice, most values $P_{SRP}(k,n)$ are positive and a typical threshold value is $\theta_{SRP} = 0.25$.

3.2. Step Two: Dilation/Erosion Process

Speech from one person mostly consists of short spurts (phonemes, words), interspersed with short silences. In obtaining a smooth speech/silence segmentation for each speaker, it is desirable to achieve two goals:

- **Goal 1**: to group spurts in order to form utterances. For a given speaker, two spurts that are separated by a small silence (e.g. less than 1 second) must be linked into the same segment.
- **Goal 2**: to remove any isolated spurt that lasts less than a minimum duration (e.g. 200 ms). We assume that such a spurt contains noise rather than speech.

Initially, we attempted to use single speaker HMMs to achieve the above goals. However, since a speech segment contains short alternating periods of speech and silence, it was found that a complex HMM topology was required, similar to that proposed for the overlaps in [3]. In addition, obtained results were significantly less than those of the previous work. In the current work, we instead achieve the above goals using an alternative approach based on simple binary dilation and erosion operators.

We apply a sequence of such operators on the binary series $ss^{(k)}$, thus achieving an effect similar to low-pass filtering in signal processing. The L-frame dilation operator for a binary sequence $u = \{u_n\}$ (with values in $\{0, 1\}$) is defined as:

$$u = \{u_n\} \quad \rightarrow \quad v = f_{dil}^{L}(u)$$

$$\text{where} \quad \forall n \quad v_n = \max\left(u_{n-L}, \ldots, u_{n+L}\right)$$

The L-frame erosion operator for a binary sequence $u = \{u_n\}$ is defined as:

$$u = \{u_n\} \quad \rightarrow \quad v = f_{ero}^{L}(u)$$

$$\text{where} \quad \forall n \quad v_n = \min\left(u_{n-L}, \ldots, u_{n+L}\right)$$

In practice, the beginning and the end of u are mirrored to solve boundary problems.

For a given speaker k, the two goals mentioned above are achieved using a succession of dilations and erosions:

$$ss^{(k)} \quad \rightarrow \quad ss2^{(k)} = f_{dil}^{L_2}\left(f_{ero}^{L_2+L_1}\left(f_{dil}^{L_1}\left(ss^{(k)}\right)\right)\right)$$

where L_1 is the maximum "small silence" duration in frames (relates to goal 1.) and L_2 is the minimum speech duration in frames (relates to goal 2.). This operation can be implemented online with a buffer of $2 \times (L_1 + L_2)$ frames, incurring a delay of $L_1 + L_2$ frames.

4. Experiments

With the two proposed methods, we segmented two data sets including segments with a single speaker and segments with two overlapping speakers. In order to compare with the single timeline of segments produced by the HMM approach [3], we combined the K binary series $ss^{(k)}$ into one sequence of integer tags (one tag per frame):

$$T_n = \sum_{k=1}^{K} ss_n^{(k)} \cdot 2^{k-1} \tag{14}$$

For each frame n, T_n describes the combination of active speakers. To assess the performance of each proposed method, we compared the sequence $\{T_n\}$ with the ground truth.

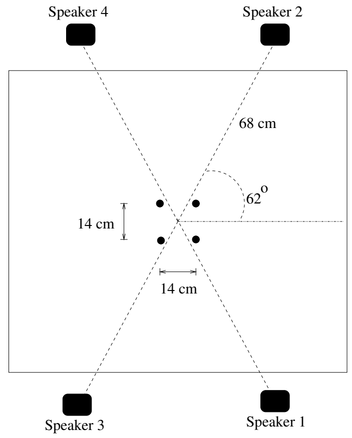

Figure 1: Experimental setup

4.1. Evaluation Criteria

To assess the system performance, we used frame accuracy (FA), precision (PRC), recall (RCL) and F-measure (F):

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

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

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

$$F \triangleq \frac{2 \times \text{PRC} \times \text{RCL}}{(\text{PRC} + \text{RCL})}$$

F 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.

4.2. Test Sets

The two test sets defined in [3] were used. Both sets were created by mixing four five-minute multichannel recordings of read speach from four speakers seated as shown in Figure 1. We used a microphone array with 4 microphones on a 14 cm-sided square.

- non-overlap test set: 9 files of 10 single-speaker segments (5 to 20 seconds per segment).
- overlap test set: 6 files of 10 single-speaker segments (5 to 17 seconds per segment) interleaved with 9 two-speaker segments (1.5 to 5 seconds per segment).

4.3. Parameters

In the experiments, we used a sampling frequency $f_s = 16kHz$ and computed features from 32 ms, 50% overlapped, Hamming-windowed frames. For the dilation/erosion process described in Section 3.2, we used $L_1 = 63$ frames (1 second) and $L_2 = 13$ frames (200 ms). We used all possible microphone pairs from the 4-element array ($P = 6$). For the SSR approach we used a diagonal matrix of ones for Σ_k (tuning it did not bring any

approach	FA	PRC	RCL	F
HMM	99.5%	1.0	1.0	1.0
SSR	99.1%	0.99	0.99	0.99
SRP	96.3%	0.85	0.96	0.90

Table 1: Results on the non-overlap test set

approach	FA	PRC	RCL	F
HMM	96.2% (88.1%)	0.93	0.93	0.93
SSR	95.4% (79.9%)	0.91	0.94	0.92
SRP	92.1% (68.9%)	0.85	0.84	0.85

Table 2: Results on the overlap test set. The FA calculated only on actual overlap segments is shown in parentheses.

significant change in the results). For the SRP approach we used $\theta_{SRP} = 0.25$.

4.4. Results and Discussion

Tables 1 and 2 show results obtained on each test set. In both sets of results, the performance of the SSR approach is comparable to that of the HMM approach, while the SRP approach performance is less but still provides a good segmentation. In particular, both approaches performed well on data containing overlapping speech. We noted that FA calculated on overlap segments was less for the two new approaches, compared to the original HMM system. This may be attributed to the fact that the new techniques do not have any explicit overlap classes, and as such do not impose any minimum duration constraint on overlap segments.

The similar performance between the HMM and SSR approaches was expectable, since exactly the same features are used in each case (see Section 3.1.1). The degradation in performance observed for the SRP approach (particularly on overlap frames) is at first surprising, since the SRP-PHAT features should be able to handle multiple concurrent speakers. Our understanding of this degradation is that it is difficult to give meaning to the absolute numerical values obtained by SRP computation. Therefore the single, constant threshold strategy defined in (13) is not an optimal approach: the true speech and silence distributions may significantly overlap and/or vary over time.

Despite this, both approaches proved effective on the read speech, including the segments with two overlapping speakers. While in these experiments we used the same data as in the previous work [3] for comparison purposes, this data does not constitute a comprehensive test-set, as it only contains read speech and is limited to overlap segments with two concurrent speakers. The proposed techniques have also been successfully applied to real meeting recordings[1] containing spontaneous speech and segments of up to four concurrent speakers, however as a ground-truth segmentation does not yet exist for these recordings, we are unable to present results at this stage.

Implicit in all of the above approaches is the assumption of prior knowledge of each speaker's location, and therefore the number of speakers. Ongoing work will investigate ways of relaxing this assumption by clustering the output of a source localisation system. Another core assumption made, is that each speaker is associated with a single region through a recording. This could potentially be addressed by combining with a speaker clustering strategy based on traditional acoustic features, such as [7].

5. Conclusion

This paper has presented two approaches for segmenting speech from multiple concurrent speakers using microphone arrays. Previous work in [3] provided a global segmentation in terms of single speaker classes and possible overlap combinations. The proposed approaches instead segment speakers individually, avoiding the need to define all possible combinations of active speakers. In this way, the major benefit of the proposed approaches is the ability to scale to all possible overlap cases (involving any arbitrary combination of speakers), with a computational load that is linear in the number of speakers. In experiments, the proposed approaches performed well on both single-speaker data and two-speaker overlap data, achieving similar performance to the global HMM strategy employed in [3]. In addition, we note that a straightforward on-line implementation is possible. Future work will verify the techniques on real meeting recordings and will aim to remove the assumption that each speaker's location is known and static throughout a recording.

[1] `http://mmm.idiap.ch`

6. Acknowledgements

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.

7. References

[1] 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*, vol. 2, 2001, pp. 1359–1362.

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

[3] G. Lathoud and I. McCowan, "Location based speaker segmentation," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, April 2003.

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

[5] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-96)*, 1996.

[6] J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments," Ph.D. dissertation, Brown University, Providence RI, USA, 2000.

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