IDIAP RESEARCH REPORT

Clustering And Segmenting Speakers And Their Locations In Meetings

Jitendra Ajmera, Guillaume Lathoud, Iain McCowan

IDIAP–RR 03-55

December 2003

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 03-55

Clustering And Segmenting Speakers And Their Locations In Meetings

Jitendra Ajmera, Guillaume Lathoud, Iain McCowan

December 2003

Abstract. This paper presents a new approach toward automatic annotation of meetings in terms of speaker identities and their locations. This is achieved by segmenting the audio recordings using two independent sources of information: magnitude spectrum analysis and sound source localization. We combine the two in an appropriate HMM framework. There are three main advantages of this approach. First, it is completely unsupervised, i.e. speaker identities and number of speakers and locations are automatically inferred. Second, it is threshold-free, i.e. the decisions are made without the need of athreshold value which generally requires an additional development dataset. The third advantage is that the joint segmentation improves over the speaker segmentation derived using only acoustic features. Experiments on a series of meetings recorded in the IDIAP Smart Meeting Room demonstrate the effectiveness of this approach.

1 INTRODUCTION

Answering a question such as "who is speaking and at what place?" is an important step toward automatic summarization of meetings. For example, based on this knowledge, a user could query a structured database to show "the last presentation made by such person" or "the last meetings attended by such person". Conversely, a user may simply want to know who attended a meeting that he missed. Globally such structuring of meeting recordings also greatly enhances the playback experience, as the user can quickly access information that is relevant to him (survey in [1]).

In the current work, we propose to segment meeting recordings in terms of *both speaker identity and speaker location*. To the best of our knowledge this has not been tried before. To achieve this, we combine ideas from two schemes that have shown good performance in past work : unsupervised mel frequency cepstral coefficient (MFCC) based speaker clustering [2] and location-based speaker segmentation [3].

The speaker clustering approach proposed in [2] is a GMM/HMM framework with minimum duration constraint. It does not use any tunable threshold/penalty term and is fully unsupervised. In past studies it has shown robustness and good performance on single channel signals, with relatively long speech segments (at least several seconds) such as broadcast recordings. However, the context here is quite different: real discussions recorded in a meeting room. Speech segments are short and many speaker changes occur, producing also overlaps. Therefore, it may well be more difficult to train accurate speaker models and obtain accurate speaker segmentation in such environment.

On the other hand, the location-based speaker segmentation proposed in [3] is able to detect speaker changes very precisely, using the sound source location as a discriminative information. It has already been successfully tested on meeting environments. In previous work, we assumed speakers in a meeting would remain at the same location, and therefore we achieved location-based speaker segmentation. But in real meetings people may stand up and move e.g. to the presentation screen. Moreover, across several meeting recordings, attendance varies and obviously the same person may be seated at different locations. So speaker identity cannot be obtained from the location information.

In this paper, we propose to exploit the complementarity of the two schemes: location information is expected to improve the speaker segmentation, thus allowing acoustic clustering to provide speaker identities. The main contributions of this paper are, first, an unsupervised approach for joint clustering and segmentation of speaker identity and location, and secondly, a simple standard deviation-based criterion for determining the number of active locations in a meeting. Preliminary results on meeting recordings show that the proposed combined approach provides promising results and improves the speaker segmentation.

Section 2 describes how acoustic and location features are extracted. Section 3 explains speaker clustering and location clustering schemes separately as well as the combined scheme. Experiments are presented and discussed in Section 4, followed by concluding remarks in Section 5.

2 FEATURE EXTRACTION

Recordings were made with a circular 8-microphone array and 4 lapel microphones. The array is used to extract the location features, while the lapel microphones are used to extract the acoustic features.

Note that in the following, "acoustic features" refers to features derived from single-channel spectral analysis such as MFCCs and LPCCs. "Location features" refers to features derived from multi-channel cross-correlation analysis. For each time frame (32ms Hamming-windowed, half-overlapping) we extract one acoustic feature vector and one location feature vector.

2.1 Acoustic Features

At each time frame, we extract 24 MFCCs (without C_0) from the 4 lapel waveforms, giving 4 concurrent streams of MFCCs. However, we need only one stream of MFCCs in order to use the algorithms described in Sections 3.1 and 3.3. Therefore, at any given time frame, we pick the MFCC vector from the lapel with the maximum energy. Preliminary experiments showed that this approach is robust to various turn-taking patterns, including overlapping speech. Results were much better compared to using MFCCs extracted from one microphone on the table. In order to avoid switching between lapels too often, we low-pass filter the energy from each lapel over consecutive frames. The whole operation is fully automatic and produces one lapel change about every 6 seconds on an average.

2.2 Location Features

We use the microphone array to locate the dominant sound source at each time frame in terms of bearing : at each time t an estimate (θ_t, ϕ_t) is produced where θ_t indicates azimuth and ϕ_t indicates elevation. This is done on all frames (speech and silence). We use a single source localization technique based on the SRP-PHAT measure [4], due to its low computational requirements and suitability for reverberant environments. For each time frame we scan a grid H of possible locations (Cartesian coordinates), select the point $\hat{Z}_t \in H$ having the maximum SRP-PHAT value, and extract spherical bearing coordinates (θ_t, ϕ_t) from $\hat{Z}_t$. The radius estimate is dropped because it is not reliable with a single microphone array. The computation of $\hat{Z}_t$ from the microphone array signals has already been presented in [5].

3 CLUSTERING APPROACH

In this section, we present the acoustic-only speaker clustering and the location clustering approaches, before presenting the proposed combined approach.

3.1 Clustering Speakers

We employ the algorithm presented in [2], which is briefly summarized in this section. The problem is formulated in an ergodic HMM framework with minimum duration constraint.

If $X = \{x_1, x_2, ..., x_T\}$ is the audio data to be segmented (described in Section 2.1), we want to find the optimal number of clusters K_S^* and their respective Gaussian mixture models (GMM) $\Lambda_{K_S^*}$ that produce the "best" segmentation of the data X according to:

$$\left(\Lambda^*_{K_S^*}, K_S^*\right) = \arg \max_{\left(\Lambda_{K_S}, K_S\right)} p\left(X, q_{best} | \Lambda_{K_S}, K_S\right) \tag{1}$$

where q_{best} is the Viterbi path with the highest likelihood. There is one state q for each speaker cluster. Thus, we want to find the set of clusters and their acoustic models that maximize the likelihood of the data; as well as the associated speaker segmentation based on this HMM topology.

The algorithm starts with over-clustering the data, i.e. clustering the data in terms of more than the expected number of classes (large initial value for K_S). This is followed by an agglomerative clustering approach where best candidate clusters are merged in an iterative fashion, trying to find a solution to Eq. 1.

In [2], we presented a merging criterion which always results in an increase of the likelihood (right hand side of Eq. 1). Summarizing the approach, if we want to decide if two clusters C_a and C_b should be merged or not, we hypothesize another cluster $C_{a \cup b}$ and model it with a number of parameters equal to the sum of the number of parameters used in modeling individual clusters C_a and C_b. Then, we compare the likelihood of the data in these two hypotheses. The important property of this approach is that it finds the optimal number of classes according to an objective function (Eq. 1) without the need of a tunable threshold/penalty term.

3.2 Clustering Locations

Using the location features described in Section 2.2, we partition the physical space in a finite set of regions $\{R_1 \cdots R_{K_L}\}$ where K_L is the number of location clusters. We assume that speakers do

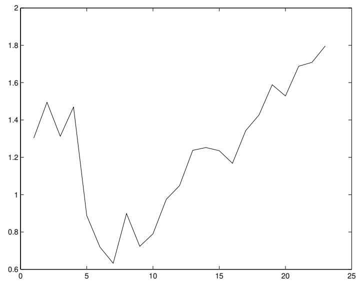

Figure 1: The value of $Q_{\{R^{(1)} \dots R^{(K_L)}\}}$ as a function of the number of location clusters. The minimum occurs at $K_L = 7$.

not move continuously from one region to another (denoted "static" assumption in the following). Practically this means a speaker can often move within a region (e.g. around a seat, or around the white-board) but rarely moves from one region to another (e.g. from a seat to the white-board).

One could use the same GMM/HMM framework as used for speaker clustering (Section 3.1). However, we are looking for a partition of the 2D space (θ, ϕ) into simple, convex regions. This leads to modeling with single Gaussian rather than GMMs: using the GMM framework and the merging criterion presented above would lead to non-convex clusters, i.e. each cluster potentially containing various, non-connected regions of the space.

Since we model each region of the space with a single Gaussian, the simplest algorithm to use is K-means, applied on the (θ_t, ϕ_t) location features. The distance metric used in the K-means is the angle between two bearings (θ_1, ϕ_1) and (θ_2, ϕ_2).

Within this framework the issue is model selection : how can we choose K_L properly? From the "static" assumption we can expect very concentrated "true" clusters in the data. We therefore define a simple standard deviation-based criterion:

$$Q_{\{R^{(1)} \dots R^{(K_L)}\}} \triangleq \sum_{k=1}^{K_L} \sigma^{(k)} \tag{2}$$

where $\sigma^{(k)}$ is the standard deviation of $\{\theta_t\}$ belonging to cluster k. If we exclude the trivial case of clusters containing only one sample, we expect that:

- when K_L is too small, at least one cluster spans over 2 or more "true" clusters, therefore leading to a large standard deviation value for that cluster. This in turn leads to a large Q value.
- when K_L is too large, the number of terms in the sum will be large so Q will be large.

Therefore, this criterion balances good fit of each cluster by a Gaussian (i.e. large K_L value) and small number of terms in the sum (i.e. small K_L value). The algorithm we implemented simply tries all values from $K_L = 1$ to a large value e.g. $K_L = 20$ and selects the K_L value yielding the partition with minimum $Q_{\{R^{(1)} \dots R^{(K_L)}\}}$ as shown in Figure 1.

The fact that we use azimuth only for the model selection criterion while we use both azimuth and elevation for the K-means distance is based upon two contradicting practical issues in our setup:

- Azimuth is expected to be the most discriminative feature considering both the horizontal planar geometry of the microphone array and the locations of the speakers.
- During silence periods, the dominant sound source is a projector, located above the table. Hence elevation is needed in the K-means distance to discriminate speakers from the projector.

3.3 Combined System

The problem is formulated as follows: if K_S is the number of speaker classes (for the acoustic stream) and K_L is the number of location classes (for the location stream), we try to segment the two streams jointly, in terms of $K_S \times K_L$ classes. In other words, we define a class for each possible active speaker at each possible location. An important point is that many of these classes may not be represented at all in the data - each speaker does not necessarily visit all locations. This is why in the following we opted for training speaker models separately from location models.

In order to achieve the joint segmentation, we propose a 2-step algorithm:

- In the first step we partition locations using the algorithm presented in Section 3.2. From the obtained partition
  $\{R_1 \cdots R_{K_L}\}$ we define a single Gaussian probability density function (pdf) as a model for each location cluster k. We use the K-means cluster centroid and cluster standard deviation to define that Gaussian pdf. These pdfs are unmodified in the next step.

- In the second step we use a modified version of the algorithm presented in Section 3.1, running in an iterative manner i.e.:
  1. Joint (speaker, location) Viterbi segmentation.
  2. Speaker models retraining.
  3. If speaker models merging is possible then merge and go back to 1. Else stop.

For the joint Viterbi segmentation, we assume independence between speaker location and speaker identity. The pdf of each state (k_{sp}, k_{loc}) is therefore:

$$p(X_{sp}, X_{loc}|k_{sp}, k_{loc}) = p(X_{sp}|k_{sp}) \cdot p(X_{loc}|k_{loc}) \tag{3}$$

where X_{sp} and $k_{sp} \in [1 \cdots K_S]$ are respectively an acoustic feature vector and a speaker state, while X_{loc} and $k_{loc} \in [1 \cdots K_L]$ are respectively a location feature vector and a location state.

4 EXPERIMENTS and Evaluation

4.1 Data

6 four-people meetings were recorded in the IDIAP Smart Meeting Room [6], each meeting lasting about 5 minutes. The four participants of each meeting were selected randomly from a set of 6 people. These meetings are part of a corpus that is fully described in [7], and can be viewed online at `http://mmm.idiap.ch`. For each meeting, people were asked to talk freely while following a short list of actions in the set { monologue, discussion, note-taking, presentation, white-board }. Participants were seated most of the time but sometimes one stood up, walked to a different location and made a presentation. 12 microphones were used: a circular 10 cm-radius 8-microphone array fixed to the table and 4 lapel microphones. We used the 8-microphone array to extract location features as explained in Section 2.2 and the 4 lapel microphones to extract acoustic features as explained in Section 2.1. The room setup is shown in Fig. 2.

To evaluate the speaker segmentation performance, a precise ground-truth (GT) segmentation of the recordings was created by an independent observer. There are 6 GT speaker clusters and 1 GT silence cluster. Each GT cluster is segmented independently in terms of "activity" and "non-activity". In other words:

- In the case of a GT speaker cluster, "active frame" means that this person is speaking - which does not exclude the possibility of other speakers also being active. Indeed, the data does have overlaps.

- In the case of the GT silence cluster, "active frame" means nobody is speaking.

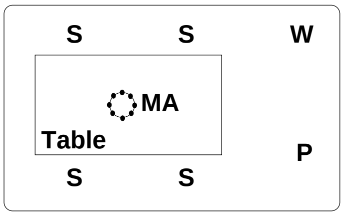

Figure 2: Recording setup. “S” denote participants’ seats, “P” presentation screen, “W” the white-board and “MA” the microphone array.

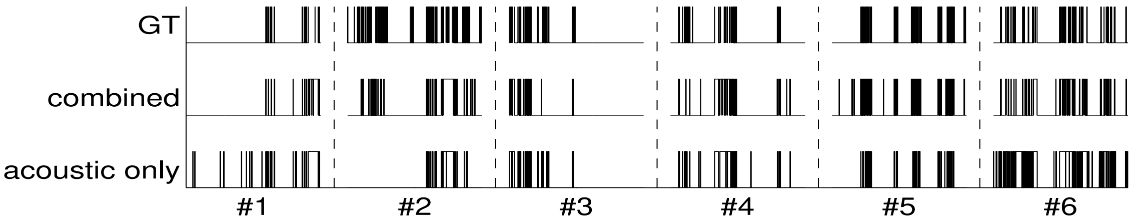

Figure 3: Time-line of speech activity for the 6 speakers. “GT” row shows the ground-truth, while the “combined” and “acoustic-only” rows show the respective results from the two systems. Each column spans over 1800 seconds of meetings.

4.2 Metrics

As explained in Section 3.3 our scheme produces a single speaker segmentation, therefore not allowing overlaps between speakers; whereas the GT does include overlaps between speakers. In order to compare with the GT, we transformed our single segmentation into a series of active/non-active segmentations, one for each speaker cluster. The data contained very short speech segments, 50% of them being shorter than 960 ms. Therefore we could not use the usual segmentation measures of precision and recall. Two metrics were used: Frame accuracy (ACC) and Half-Total Error Rate (HTER). ACC is the overall proportion of correctly classified frames. HTER is the average of False Alarm Rate (FAR) and False Rejection Rate (FRR). FAR is the proportion of erroneous frames in the active frames of the result. FRR is the proportion of erroneous frames in the non-active frames of the result.

We used HTER to determine the best combinatorial match between GT speaker identities and result speaker identities.

4.3 Results

We first ran the acoustic clustering algorithm described in Section 3.1 alone, then the combined system. While both systems provided the correct number of acoustic clusters (6 speakers and 1 silence), the quality of the speaker segmentation was improved in terms of both HTER and ACC, by using the combined system, as shown in Table 1. Calulations show that the improvement in ACC is statistically significant.

System	HTER	ACC
Acoustic Only clustering	19.2	92.6
Combined Clustering	17.3	94.6

Table 1: Speaker segmentation performance. HTER and frame accuracy (ACC) are expressed as percentages.

The actual speaker segmentation results and ground-truth are shown in Fig. 3. It shows the entire speech/silence time-line for each speaker. This high-level view shows the concatenation of the 6 meetings, in other words it shows how usable the results are to answer the question "who attended that meeting?". Improvement brought by the combined system is clearly visible for speakers #1 and #2, while results for the other speakers are similar to the acoustic-only results.

We also looked at the locations determined by the K-means clustering in the combined system. As shown in Fig. 1, the algorithm chose to partition the space into $K_L = 7$ regions. The centroid values given by the K-means algorithm corresponded to the 6 main speaker locations shown in Fig. 2 plus the projector. The projector cluster was expected as it is a dominant source of energy during silence.

5 CONCLUSION

We have proposed an unsupervised approach for segmenting meeting recordings jointly in terms of speaker identity and speaker location. Such a segmentation is important in the context of browsing or searching a meeting corpus. The proposed combined approach is fully unsupervised, and free of any tunable threshold. The main achievement of this work was to automatically infer the number of clusters for both speaker and locations as well as the joint segmentation. In our experiments, we observed that the proposed combined approach also improves speaker segmentation, as compared with the acoustic only clustering. Experiments were carried out on a set of meetings recorded with both close-talking and distant microphones.

Future work will explore the analysis of concurrent speakers: overlapping speech occurs regularly in meetings. It was shown in previous work that location-based analysis has a strong potential for this. A second direction is toward continuous tracking, as opposed to the discrete partition of locations used here. Finally, more comprehensive testing will be undertaken to further illustrate the relevance and robustness of our approach.

6 Acknowledgments

The authors acknowledge the support of the Swiss National Science Fund through the MULTI project and European Union through the M4 and HOARSE projects. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2.

References

[1] R. Cutler, Y. Rui, A. Gupta, J.J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings : A meeting capture and broadcasting system," in *Proc. of ACM Multimedia*, 2002.

[2] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proc. of ASRU*, 2003.

[3] G. Lathoud and I. McCowan, "Location based speaker segmentation," in *Proc. of ICASSP*, 2003.

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

[5] D. Gatica-Perez, G. Lathoud, I. McCowan, and J.-M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," in *Proc. of ICCV-WOMTEC*, 2003.

[6] D. Moore, "The IDIAP Smart Meeting Room," IDIAP-COM 07, IDIAP, 2002.

[7] I.A. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interactions in meetings," in *Proc. of ICASSP*, 2003.