Multi-Stream Speaker Diarization Systems for the Meetings Domain

Ascensión Gallardo-Antolín[1], *Xavier Anguera*[2] *and Chuck Wooters*[2]

[1]Department of Signal Theory and Communications
Carlos III University of Madrid, Leganés (Madrid), Spain
[2]International Computer Science Institute (ICSI)
1947 Center St., Suite 600, Berkeley, CA 94704, USA
gallardo@tsc.uc3m.es, {xanguera, wooters}@icsi.berkeley.edu

Abstract

In the context of speech and speaker recognition systems, it is well known that the combination of different feature streams can improve significantly their performance. However, the application of multi-stream (MS) techniques to speaker diarization systems has not been extensively studied. In this paper, we address this issue: we formulate different MS techniques, such as feature combination, probability combination and selection, for their specific application to the segmentation and clustering modules of a speaker diarization system. We evaluate the different methods proposed for the meetings domain (RT04s database) and two different pairs of streams: first, MFCC and PLP and second, MFCC and prosodic features. For both types of multi-streams, results show that the MS probability combination approach applied to the segmentation stage clearly outperforms the single-stream, MS feature combination and MS selection systems.

Index Terms: speaker diarization, multi-stream features, prosodic features.

1. Introduction

Speaker diarization is the task of dividing an input audio recording into speaker-homogeneous regions and providing the same label to the segments uttered by the same speaker. Usually, it consists of two different stages: segmentation, in which the speaker changes are located, and clustering, in which segments corresponding to the same speaker are grouped. One of the main difficulties of speaker diarization is that it has to be treated as an unsupervised problem because there is no prior information about the number of speakers, their identity or the acoustic conditions [1], [2].

Speaker diarization has been extensively studied in the Broadcast News (BN) domain and more recently, in the meetings domain [3]. These environments present several differences: speech in meetings is more spontaneous than in BN and is distorted because of the use of distant microphones. In addition, cross-talk is more frequent than in BN [2]. In this paper, we have focused on the meetings domain.

One of the issues in speaker diarization systems is the choice of the best acoustic representation of the audio signal. Different types of acoustic features have been considered in the literature. Most of them were initially devised for speaker recognition, such as: LSP [4], LPCC [5], PLP [6], MFCC [5], [7] with their first [1] and second derivatives [2].

In the context of speech technologies, it is well known that the combined use of different acoustic representations with complementary information can improve significantly system performance. Concerning this point, two main question arise: (1) Which is the more effective way of incorporating information from different feature streams? and (2) In which stage(s) of the system is it preferable to combine them?.

In recent years, in the field of ASR, three Multi-Stream (MS) approaches have been proposed: namely, Feature Combination (MS-FC), Probability Combination (MS-PC) [8] and Selection (MS-S) [9]. In the first alternative, a single vector is built as the concatenation of several types of acoustic features, so the acoustic modelling and decoding processes are performed in the same way as in the single-stream case. In FC, usually in order to reduce the high dimensionality of the resulting vectors, a Principal Component Analysis (PCA) is applied. In the second and third approaches, acoustic models are built separately for each feature stream and their probabilities are somehow jointly used in the decoding stage: for MS-PC they are combined according to some predefined rule and for MS-S, only the stream with highest probability is dynamically selected. However, in the context of speaker diarization systems, multi-stream techniques have not been intensively studied.

With respect to the second question, as the main processes involved in speaker diarization (segmentation and clustering) pursue different goals, it would be possible that features which are effective for segmentation may not be appropriate for clustering and viceversa. This issue has been addressed (at least, indirectly) in different studies, generally for the BN environment, although no definitive conclusions have been made. For example, in [11], 13 MFCC with first and second derivatives were used for segmentation and 13 PLP with c0 for clustering. Another relevant example is [10], in which 45-dimensional vectors composed of PMVDR ("Perceptual Minimum Variance Distortionless Response"), FBLC ("Filterbank Log Coefficients") and SZCR ("Smoothed Zero Crossing Rate") were used in the segmentation of BN, whereas a feature combination of PMDVR and FBLC was used for a false alarm compensation procedure (which, conceptually, is similar to a clustering applied over two adjacent segments).

In this paper, we have addressed the two issues mentioned before for speaker diarization in the meetings domain. We have reformulated (if needed) the multi-stream strategies for their application to the segmentation and clustering modules and we have tested the proposed approach for two different pairs of streams: first, MFCC

and PLP and second, MFCC and prosodic features.

The remainder of the paper is organized as follows. In Section II we describe the baseline system. In Section III we introduce different alternatives for multi-stream speaker diarization systems. Section IV describes the experimental results. Finally, in Section V we present our conclusions.

2. System description

For our experimentation we have used the ICSI-SRI Spring 2005 diarization system [7]. As can be observed in Figure 1, it is mainly composed of two modules: the data preprocessing module and the diarization system itself.

In the preprocessing module, the audio signal is passed through a speech activity detector in order to detect and discard the non-speech portions. Then, the speech regions are coded into acoustic parameters. Some details about the parameterization process and the type of features used will be described in section 4.

The speaker diarization process is an iterative segmentation-clustering algorithm which uses a measure based on the Bayesian Information Criterion (BIC) as stopping criterion.

The initialization of the algorithm divides the audio data into K clusters of equal length and builds an initial GMM speaker model for each cluster. Next, the audio data is re-segmented using a Viterbi decoder which attempts to find the optimum sequence of models (speakers) according to the Maximum Likelihood (ML) criterion. Then, the GMM model of each cluster is re-trained with the segments assigned to it in the previous step. The next sub-module consists of an agglomerative clustering in which a modified version of the conventional BIC is applied in order to select the pair of clusters to be grouped (if any). After the merge of this pair of clusters, their corresponding models are combined into one single model. The whole process is repeated until no pair of clusters meets the merge criterion.

The variation in BIC introduced in [5] avoids the existence of tunable parameters (as, for example, the penalty factor in the conventional BIC). For the purpose of this paper, this is a desirable characteristic of the diarization system because it allows the experimentation with different types of features without the need of adjusting threshold values each time a new feature stream is used.

3. Using multiple feature streams in diarization systems

The goal of the segmentation and clustering stages are different from one another, and therefore, features which are effective for segmentation may not be appropriate for clustering. Features used for segmentation attempt to discriminate between segments with different acoustic conditions (speaker, environment, channel, ...), i.e, maximize the difference between segments with any kind of acoustic difference in order to properly place the changing points, whereas in speaker clustering, features attempt to merge segments containing a single speaker in spite of the background acoustic conditions or any kind of distortions (noise, distance to microphone, etc). In addition, segmentation and clustering have different objective functions to maximize: ML and BIC, respectively.

So, in order to incorporate information from one or more feature streams more efficiently, it would be interesting to consider separately each of these stages. In this section we address this issue by means of the formulation and application of two different multi-stream approaches (combination and selection) to the segmentation and clustering stages of the diarization system.

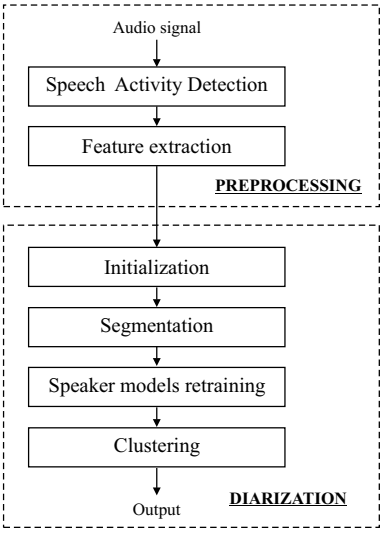

Figure 1: *Block diagram of the speaker diarization system.*

3.1. Alternatives in the segmentation stage

3.1.1. Multi-Stream Probability Combination (MS-PC)

Given N feature streams $x = \{x_1, \ldots, x_n, \ldots, x_N\}$ and K classes (clusters) $\mathbb{C} = \{c_1, \ldots, c_K\}$, their individual probabilities can be combined at frame level via the (weighted) product rule following this expression,

$$p(x_1, \ldots, x_n, \ldots, x_N|c_k) = \prod_{n=1}^{N} [\bar{p}(x_n|c_k)]^{\gamma_n} \quad (1)$$

where $p(x_1, \ldots, x_n, \ldots, x_N|c_k)$ is the combined likelihood used in the Viterbi decoder, γ_n are the stream weights and $\bar{p}(x_n|c_k)$ is the normalized likelihood of the feature stream vector x_n given the class c_k and its computed as follows,

$$\bar{p}(x_n|c_k) = \frac{p(x_n|c_k)}{\sum_{k=1}^{K} p(x_n|c_k)} \quad (2)$$

This normalization is necessary because of the different ranges of the individual likelihoods $p(x_n|c_k)$. Although, other normalizations can be applied [12], we have empirically determined that Equation (2) achieves the best performance in our system.

3.1.2. Multi-Stream Selection (MS-S)

Given a single-stream acoustic representation x and a set of K classes $\mathbb{C} = \{c_1, \ldots, c_K\}$, the purpose of the Viterbi decoder is to find the optimum class sequence C^* that maximizes the probability of the class sequence given the acoustic observation, x.

$$C^* = \arg\max_{C} p(C|x) \quad (3)$$

For example, in Figure 2, if we consider each feature stream S_1 and S_2 individually, the optimal sequence is $\{c_2, c_0, c_2\}$ when using S_1 and it is $\{c_1, c_3, c_2\}$ for S_2.

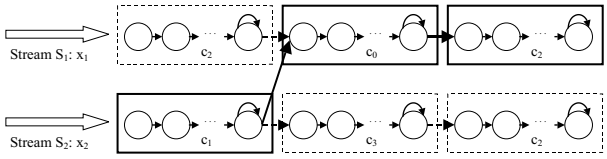

Figure 2: *Viterbi decoding in the MS-Selection approach.*

However, when multiple streams with complementary information are used, the optimum path would be the concatenation of the best partial paths produced by any of the individual streams. For example, in Figure 2, the overall class sequence is $\{c_1, c_0, c_2\}$.

Therefore, if x is the set of N streams $x = \{x_1, \ldots, x_n, \ldots, x_N\}$, it is possible to search the best feature stream in addition to search the best class sequence using a modified version of the conventional Viterbi algorithm [9], in which the objective is now:

$$C^* = \arg\max_{C}\{\max_{n \in N} p(C|x_n)\} \quad (4)$$

This is the basic formulation of the MS-Selection approach for the segmentation stage, which, in practice, selects the best feature stream and the best class according to the ML criterion. Note that the posterior probabilities $p(C|x_n)$ involved in Equation (4) have to be rewritten in terms of the acoustic likelihoods $p(x_n|C)$ and they need to be normalized as in subsection 3.1.1.

As suggested in [9], Equation (4) only is used to determine the best feature stream for each decision boundary, which corresponds to the possible transitions from one speaker to another. Also, a switching factor for each stream (denoted by γ_n) is introduced in order to control the transition probability between streams (and, therefore, the contribution of each stream to the final sequence).

3.2. Alternatives in the clustering stage

3.2.1. Multi-Stream Probability Combination (MS-PC)

The likelihoods involved in the BIC computation are calculated using the weighted rule product mentioned in subsection 3.1.1 (Equations (1) and (2)).

3.2.2. Multi-Stream Selection (MS-S)

The idea is to simultaneously decide which pair of clusters is going to be merged and which feature stream is going to guide this process according to the BIC criterion. The algorithm consists of the following steps:

- Step 1. For each feature stream, compute all the BIC-values between all the clusters and determine the pair of clusters, that is most likely to be merged following the BIC criterion. This way, we have several candidates to be merged.
- Step 2. Merge the pair of clusters with a highest positive BIC-value.

Again, the probabilities involved in the computation of BIC must be normalized as in subsection 3.1.1.

4. Experimental results

The data used in our experimentation consist of the RT04s development and evaluation sets used in the NIST Rich Transcription 2004 evaluations for the meetings domain [13]. Each set is composed of 8 excerpts of meetings of approximately 10 minutes long. The development set was used to determine the optimum stream weights in the experiments described in next subsections. In this paper, we have considered the SDM (Single Distant Microphone) condition defined by NIST, which uses only the audio data recorded with the most centrally located microphone.

The performance is measured in terms of Diarization Error Rate (DER) which is the percentage of time that the system assigns a wrong speaker label. More details can be found in [13].

All the experiments reported in this paper have used the same configuration parameters in the diarization system: the number of initial clusters was 10, the GMM models were initially composed of 5 gaussians and the minimum segment length was 3 s.

4.1. Results with MFCC and PLP parameterizations

In this section, we describe the results obtained with the different multi-stream approaches proposed when considering two streams of acoustic features: MFCC and PLP parameters. In both cases, the audio signal was analyzed with a Hamming window of 30 ms long and 19 (MFCC or PLP) coefficients were extracted at a frame period of 10 ms. Note that we have used parameters of a higher order than in ASR tasks (usually, 12 MFCCs or 12 PLPs) in order to incorporate more speaker-dependent information.

Table 1 shows the Diarization Error Rates (%) on the development and evaluation sets of the RT04s database, and the values of the stream weights (γ_1 for MFCC and γ_2 for PLP) used.

The first and second row contain the results attained with the single streams. As can be observed, PLP outperforms MFCC on the *dev04s* set, but it is significantly worse than MFCC on the *eval04s* set. For comparison purposes, we have included in the third row the result for the MFCC-PLP feature combination with PCA (23 components). In this case, the DER was increased on both sets with respect to the DERs obtained with the individual streams. This result suggests that feature combination is not a good choice in our system.

The fourth and fifth rows correspond to the DERs obtained with the multi-stream selection (MS-S) and combination (MS-PC) approaches, respectively, in the clustering stage. In this case, only MFCC were used for segmentation. MS-S achieves a worse performance than the two single streams. We think that a maximization of the BIC value along different feature streams does not predict correctly which pair of clusters is more likely to be merged as long as not all the feature streams are equally "good" for performing this operation. With respect to MS-PC, its DER is slightly worse than PLP and only outperforms the result on the *dev04s* in the case of MFCC. In short, we can conclude that the use of multi-stream approaches in the clustering stage does not produce any improvement with respect to the single-stream systems.

The two last rows contain the DERs corresponding to MS-S and MS-PC approaches, respectively, in the segmentation stage. In this case, clustering was performed using only MFCC. As can be observed, MS-S achieves better results than the two single-stream systems; in fact, it is capable of slightly outperforming the best single-stream result on each set: for example, in comparison with MFCC, the DER on *eval04s* is reduced by almost 0.5% absolute (from 16.38% to 15.91%), but in comparison with PLP, the DER present a much greater drop of around 2.3% (from 18.18% to 15.91%). Finally, the best performance is obtained with the MS-PC approach in segmentation, which achieves a relative error reduction of 8.17% and 4.23% with respect to MFCC and PLP, respectively, on *dev04s* and a relative reduction of 6.04% and

15.35% with respect to MFCC and PLP, respectively, on *eval04s*.

In conclusion, it appears to be more effective to combine multiple streams in the segmentation stage than in the clustering stage. One possible explanation is that features studied here are more suitable for segmenting than for grouping.

Stage	Features	DEV04s	EVAL04s
Both	MFCC	19.46%	16.38%
Both	PLP	18.66%	18.18%
Both	FC: MFCC+PLP	20.71%	19.02%
Clustering	MS-Selection	22.69%	20.09%
Clustering	MS-P. Combination ($\gamma_1 = 0.9; \gamma_2 = 0.1$)	18.97%	18.27%
Segmentation	MS-Selection ($\gamma_1 = 0.8; \gamma_2 = 0.2$)	18.35%	15.91%
Segmentation	MS-P. Combination ($\gamma_1 = 0.9; \gamma_2 = 0.1$)	17.87%	15.39%

Table 1: *Results for different alternatives of the multi-stream diarization system with MFCC and PLP parameterizations.*

4.2. Results with MFCC and prosodic features

Recent studies show that the use of prosodic information can be very useful for distinguishing between speakers. Therefore, we decided to include pitch-related features as an input in the diarization system in combination with MFCC.

The pitch was extracted from the audio signals using the Snack program [14] every 10 ms with an analysis window of 40 ms. It was computed only in voiced regions and an interpolated value of it was assigned to unvoiced regions. In order to avoid possible pitch tracker irregularities, F0 sequences were filtered out using a median filter with a window of 5 frames. Then, the logarithm of F0 and its first derivative were calculated.

Table 2 shows the DERs obtained on the RT04s database for MFCC, a feature combination (FC) of MFCC and prosodic features with PCA (19 components) and the multi-stream combination approach (MS-PC) in the segmentation stage. As can be observed, FC increases the DER with respect to MFCC by almost 5.8% absolute on *dev04s* and by 6.8% on *eval04s*. This result suggests that special care must be taking in incorporating prosodic information in the diarization system. However, MS-PC clearly outperforms the single-stream system: in fact, it achieves a relative DER reduction of 11.46% and 3.79% with respect to MFCC on *dev04s* and *eval04s*, respectively. Therefore, it seems that the MS approach better exploits the complementarity between features of very different nature for speaker change detection.

Stage	Features	DEV04s	EVAL04s
Both	MFCC	19.46%	16.38%
Both	FC: MFCC+ +log-F0+Δlog-F0	25.21%	23.11%
Segmentation	MS-P. Combination ($\gamma_1 = 0.9; \gamma_2 = 0.1$)	17.23%	15.76%

Table 2: *Results for different alternatives of the multi-stream diarization system with MFCC and prosodic features.*

5. Conclusions

In this paper, we have presented the use of multi-stream processing techniques in speaker diarization systems. Experiments with MFCC and PLP streams showed that the single-stream approach is more suitable for the clustering stage, whereas the use of multi-stream combination (MS-PC) in the segmentation module clearly outperforms other techniques: single-stream, feature combination and selection. These results have been confirmed when using MFCC and prosodic information as feature streams. In this case, MS-PC achieves a relative DER reduction of 11.46% and 3.79% with respect to MFCC on the RT04s development and evaluation sets, respectively. In future work, we plan to explore combinations of more than two feature streams, as for example, MFCC, PLP and prosodic features.

6. Acknowledgements

The authors would like to thank Nikki Mirghafori, James Fung, Barbara Peskin and Juan M. Montero for their help and useful comments.

7. References

[1] Zhou, B. and Hansen, J. H. L., "Efficient audio stream segmentation via the combined T^2 statistic and bayesian information criterion", *IEEE Trans. Speech and Audio Proc.*, 13(4):467–474, July 2005.

[2] Meignier, S., Moraru D., Fredouille, C., Bonastre, J.-F., and Besacier, L., "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer, Speech and Language*, 20, pp. 303-330, 2006.

[3] Reynolds, D. A. and Torres-Carrasquillo, P., "Approaches and applications of audio diarization," *ICASSP'05*, 2005.

[4] Adami, A, Kajarekar, S. and Hermansky, H., "A new speaker change detection method for two-speaker segmentation," *ICASSP'02*, Orlando, USA, May 2002.

[5] Ajmera, J. and Wooters, C., "A robust speaker clustering algorithm," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'03)*, Virgin Islands, USA, 2003.

[6] Sinha, R., Tranter, S. E., Gales, M. J. F. and Woodland, P. C., "The Cambridge University March 2005 speaker diarisation system", *Interspeech'05*, pp. 2437-2440, Lisbon, 2005.

[7] Anguera, X., Wooters, C., Peskin, B. and Aguiló, M., "Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System," *Proc. of NIST MLMI Meeting Recognition Workshop*, Edinburgh, 2005.

[8] Ellis, D. and Bilmes, J., "Using mutual information to design feature combinations," *ICSLP'00*, Beijing, China, 2000.

[9] Jiang, L. and Huang, X., "Unified decoding and feature representation for improved speech recognition," *EUROSPEECH'99*, Budapest, Hungary, September 1999.

[10] Huang, R. and Hansen, J. H. L., "Advances in unsupervised audio segmentation for the Broadcast News and NGSW corpora," *ICASSP'04*, Montreal, Canada, May 2004.

[11] Tranter, S. E., Gales, M. J. F., Sinha, R., Umesh, S. and Woodland, P. C., "The development of the Cambridge University RT-04 diarisation system," *Fall 2004 Rich Transcription Workshop (RT04)*, Palisades, NY, November 2004.

[12] Kirchhoff, K., Fink, G. and Sagerer, G., "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, 37, pp. 303-319, 2002.

[13] National Institute for Standards and Technology. NIST rich transcription evaluations, website: http://www.nist.gov/speech/tests/rt.

[14] "The Snack Sound Toolkit", KTH, website: http://www.speech.kth.se/snack/.