Robust Speaker Diarization for Meetings: ICSI RT06s evaluation system

Xavier Anguera[1,2]*, Chuck Wooters*[1]*, Jose M. Pardo*[3]

[1] International Computer Science Institute
1947 Center St., Suite 600
Berkeley, CA 94704, U.S.A.
[2] Technical University of Catalonia (UPC)
08034 Barcelona, Spain
[3] Universidad Politecnica de Madrid
28040 Madrid, Spain
{xanguera,wooters, jpardo}@icsi.berkeley.edu

Abstract

In this paper we present the ICSI speaker diarization system submitted for the NIST Rich Transcription evaluation (RT06s) [1] conducted on the meetings environment. This is a set of yearly evaluations which in the last two years have included the speaker diarization of two kinds of distinct meetings: the conference room and the lecture room. The system presented focuses on being robust to changes in the meeting conditions by not using any training data. In this paper we introduce four of the main improvements to the system from last years' submission: The first is a new training-free speech/non-speech detection algorithm. The second is the introduction of a new algorithm for the system initialization. The third is the use of a frame purification algorithm to increase clusters differentiability. The last improvement is the use of inter-channel delays as features, greatly improving performance. We show the diarization error rate (DER) score of this system on all available meetings datasets to date for the multiple distant microphone (MDM) and single distant microphone (SDM) conditions.

1. Introduction

The goal of speaker diarization is to segment an audio recording into speaker-homogeneous regions [2] answering the question "Who spoke when?". Typically, this segmentation must be performed with little knowledge of the characteristics of the audio or of the participants in the recording. We normally know the type and source of the recording (wether it is a meeting or broadcasted news, and when/where it happened). We cannot use any information on the number of speakers present or their identities, and where there is noise, commercials or other events.

Probably the single mostly-used technique in speaker diarization is agglomerative clustering. An initial set of clusters is iteratively reduced by merging the closest pair according to a similarity metric until a stopping point is reached. A cluster is defined to be a set of segments not necessarily contiguous that share some acoustic similarity. A segment is defined to be a contiguous set of acoustic frames. In the system presented here we constrain the segments to have a minimum duration. It is also common practice to use the Bayesian Information criterion (BIC) as a similarity metric between clusters and as a stopping criterion.

This is the forth year that NIST has put together a rich transcription evaluation for the meetings environment, and the second year that speaker diarization is one of the proposed tasks. In both speaker diarization evaluations there has been a distinction between the "conference room" meetings and the "lecture room meetings". The first are highly interactive meetings around a conference room table. Examples of these meetings are the ICSI Meetings project, the AMI project and meetings recorded at NIST, CMU, Virginia Tech(VT) and others. The second type are mainly less interactive lectures where a presenter talks standing up in a designated area. An example of these is the CHIL project recordings.

For the RT06s evaluation, ICSI submitted four systems for the multiple distant microphones (MDM) condition and two systems for the single distant microphone (SDM) condition in the conference room environment, and four systems for MDM, two systems for SDM and four systems for the all-distant microphones (ADM) condition in the lecture room environment. In this paper we focus on the primary systems submitted for MDM and SDM in the conference room, although some of the changes were also applied to the other systems when appropriate.

In next section we review the general blocks on which the MDM system is based, sections 3 through 6 introduce the main changes in the system from the last submission in RT05s and section 7 shows the results of this year's system on the datasets from all evaluations to date.

2. Agglomerative Speaker Diarization System

As explained in [3], the speaker clustering system is based on an agglomerative clustering technique. Its main blocks are shown in figure 1. It initially splits the data into K clusters (where K must be greater than the number of speakers and is chosen using the algorithm presented in [4]), and then iteratively merges the clusters (according to a merge metric based on ΔBIC) until a stopping criterion is met. Our clustering algorithm models the acoustic data using an ergodic hidden Markov model (HMM), where the initial number of states is equal to the initial number of clusters (K). Upon completion of the algorithm's execution, each remaining state is taken to represent a different speaker. Each state in the HMM model contains a set of MD sub-states, imposing a minimum duration on the model (we use $MD = 3$ seconds). Within the state, each one of the sub-states shares a probability density function (PDF) modelled via a Gaussian mixture model (GMM).

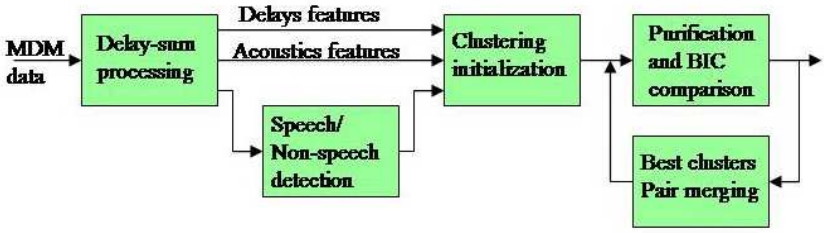

Figure 1: *RT06s Speaker Diarization system blocks diagram*

The system works as follows:

1. If more than one recorded channel is available for a given meeting recording, combine them all into a single "enhanced" channel using a delay-and-sum algorithm further described in [5].
2. Run a speech/non-speech detection on the input data using the speech/non-speech algorithm presented in [6] and explained in section 3.
3. Extract acoustic features from the data and remove non-speech frames from the agglomerative processing.
4. Estimate the number of initial clusters K using the algorithm presented in [4].
5. Create models for the K initial clusters using a new clusters initialization algorithm explained in section 4.
   (a) Run a Viterbi decode to resegment the data.
   (b) Retrain the models using the Expectation-Maximization (EM) algorithm and the segmentation from step (a). Iterate between (a) and (b) until the segmentation stabilizes.
   (c) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.0 using the frame purification technique introduced in [7] and section 5.
   (d) If no such pair of clusters is found, stop and output the current clustering.
   (e) Merge the pair of clusters found in step (c). The models for the individual clusters in the pair are replaced by a single, combined model.
   (f) Go to step (a).

For stopping criteria of the merging and clustering, we use a variation of the commonly-used BIC [8]. The ΔBIC compares two possible models for two clusters: belonging to the same speaker or to different speakers. The variation used was introduced by Ajmera et al. [9], and consists of the elimination of the tunable parameter λ by ensuring that, for any given ΔBIC comparison, the difference between the number of free parameters in both models is zero.

One of the main overall changes for this year is that we eliminated all remaining dependency of our system to training data. This was achieved by the creation of a training-free speech/non-speech detector introduced in next section. Furthermore, this year we introduced the use of data other than the acoustic data for the clustering by successfully using the delays between channels (in the MDM condition) as a new feature stream in the agglomerative clustering. This is further explained in section 6. Apart from these, a new clustering initialization algorithm and a frame purification algorithm contributed to the increase in the system's robustness and therefore improved its performance. The following sections introduce all these techniques. Finally, section 7 shows experiments on several meetings datasets and then some conclusions are drawn.

3. Speech/Non-Speech Detection Algorithm

When doing speaker diarization, the less non-speech that we allow into the agglomerative clustering the better that the system is going to be at correctly finding the optimum amount of speaker and their optimum clustering. Until the RT05s evaluation, the speech/non-speech system we were using was based on pre-trained acoustic models for both speech and non-speech, using pre-labelled training data. This forced the retraining of the models every time new meetings environments were put forth, e.g. "conference room" versus "lecture room" data. For this year's evaluation we have developed a new speech/non-speech detector that is training-free and therefore more robust to unseen meetings data, as long as the main non-speech event in the recording is silence.

The created system is a hybrid energy-based detector and model-based decoder. In the first stage an energy-based detector finds all segments with low energy throughout the recorded meeting. A variable threshold adapts automatically to obtain enough non-speech segments, and moves on to the second module. At this second stage, speech and non-speech models are trained using the segmentation from the first stage and then several cycles of Viterbi segmentation and model retraining take place, finally outputting the speech/non-speech segmentation. The total speech/non-speech detection error this year is similar to that of the prior system, but when used in the speaker diarization system, we obtain significant improvements in DER.

4. Cluster Initialization Algorithm

In order for the agglomerative clustering to work properly in obtaining the optimum number of clusters for a particular recording, we need to initialize the system with N (being $N > N_{opt}$ the optimum amount of clusters) clusters containing acoustically homogeneous data from only one speaker.

For this purpose there was the belief that any initialization of the clusters would be able to perform well provided that the created models were iterated a few times, resegmenting the data and retraining them to allow for all acoustically homogeneous data to come together. With this is mind, last year's ICSI Speaker Diarization system was using an initial clustering where each cluster was trained with a segment of contiguos acoustic frames evenly dividing the data into N segments. Although being a very simple technique and working extremely well for some cases, in many others

the resulting clusters would contain more than one speaker which would affect the (5c-d above) stopping criterion causing the final DER to increase. This is believed to be one of the causal factors of an increase in per-show flakiness, as defined in [10].

The new initialization algorithm, explained in [11], consists of 3 stages of processing. First, a speaker-change detection using the Bayesian Information Criterion (BIC) metric is used to define acoustically similar segments. The second stage compares these segments and creates groups of segments (friends) which are as different as possible from each other group (enemies) according to a normalized cross-likelihood metric. Once N groups are defined, their models are created and a segmentation is performed to distribute all the data among all models. Using this technique, we obtain an increase in cluster purity right after the initialization process and a general improvement of the overall DER, together with a more robust system.

5. Frame Purification for Clusters Comparison

By using an agglomerative clustering technique to obtain the optimum amount of final clustering, the system's performance heavily relies on the metric used to compare the similarity between clusters pairs as well as the clustering stopping criterion. Non-speech data is one of the main causal factors of anomalous behavior, which is the reason a speech/non-speech detector is being used prior to the clustering process. The data considered to be speech still contains small non-speech segments (normally silence segments in the meetings environment) and other unvoiced speech which impedes the appropriate differentiability between clusters.

The frame purification algorithm (explained in [7]) detects and eliminates such acoustic frames from affecting the cluster models during the BIC comparison. To do so, it uses a metric related to the likelihood of the frames given the acoustic model. When the cluster models' complexity is greater than 2 gaussian mixtures it is shown that non-speech frames always obtain the highest likelihoods, indicating that these are modelled by narrower gaussian mixtures. An important improvement on the clusters' differentiability is obtained by filtering out such frames. This method is demonstrated to work better than filtering based on average frame energy.

6. Use of Inter-Channel Delays in Clustering

Possibly this years most noticeable improvement is the inclusion of the inter-channel delays for the tasks where more than one microphone is available for processing. Such delays are the result of the delay-and-sum analysis of all input signals that results in the creation of an enhanced signal from multiple channels. For inclusion in the clustering, delays are computed between a reference channel and all other channels at the same rate as the acoustic features. The delays are modelled using single gaussian mixtures, with the same minimum duration as the acoustic features and are used to represent the same speaker segments as their acoustic counterpart. When two clusters merge, their delay models are combined in the same way as the other models.

Both the delay models and the acoustic models are used to classify the data into the different clusters via a Viterbi segmentation and for clusters comparison using BIC. In both cases the joint likelihood for any given frame is computed as:

$$lkld(x[n]|\Theta_{aco},\Theta_{del}) = \alpha \cdot lkld(x[n]|\Theta_{acc}) + (1-\alpha)\cdot lkld(x[n]|\Theta_{del}) \quad (1)$$

Where Θ_{aco} is the acoustic model, Θ_{del} is the delay model and α weights the effect of each model in the system. The value for α needs to be optimized using development data, usually obtaining the optimum values for $\alpha \sim 0.9$.

7. Experiments

In order to test this year's diarization system on the meeting diarization task we use all datasets provided by NIST for meetings. These are from the evaluations RT02s, RT04s, RT05s and RT06s. From the evaluations in RT05s and RT06s we constrained our test to the conference room meetings, which have similar characteristics to those in the previous years. The conference room meetings consist on excerpts of various length (depending on the year) recorded by various institutions (NIST, LDC, AMI, ICSI, etc). The available number of microphones and the room setups also varies from meeting to meeting, and ranges from 1 microphone in the RT02s and RT04s CMU recordings to 16 microphones in the RT05s and RT06s AMI meetings. Both RT02s and RT04s contain 8 meetings, RT05s has 10 meetings and RT06s has 9, of which only 8 were finally scored.

In order to evaluate diarization performance we make use of the Diarization Error Rate score used by NIST in the RT evaluations. It is computed by first finding an optimum one-to-one mapping of reference speaker ID to system output ID and then obtaining the error as the percentage of time that the system assigns the wrong speaker label. In the RT06s evaluation, the primary evaluated metric includes overlap regions, where more than one speaker talks at the same time, and therefore a missed-speech error is reported if both speakers are not detected.

In all evaluations to date the reference segmentations used to evaluate the system performance have been created by hand-listening to the individual head-mounted microphones (IHM) data. In table 1 we show the DER for MDM and SDM considering both overlap (main metric in RT06s) and non overlap (main metric in previous years).

Evaluation	SDM		MDM	
campaign	non ovl.	ovl.	non ovl.	ovl.
RT02s	18.91%	25.05%	20.79%	26.95%
RT04s	14.29%	31.23%	15.44%	30.55%
RT05s	15.04%	23.46%	**10.41%**	18.73%
RT06s	31.25%	43.56%	23.06%	36.99%

Table 1: System DER on different evaluation campaigns using hand-made references

Comparing this year's system with last year's results on the same condition (MDM, non overlap), we obtain an improvement of 44% relative of 10.41% versus 18.56% obtained with the RT05s evaluation system.

During the RT06s evaluation, there have been rising concerns about the suitability of hand-made references, given their generation cost in manpower and their accuracy. In addition, we have observed a lack of consistency in the reference segmentations between different years' datasets, together with the existence of varying quantities of extra padding for the overlap speaker turns, which cause varying extra amounts of missed-speech errors. In general, we believe that the hand made speaker segmentation references

are somewhat valid to compare different diarization systems when using them on the same data, but they show too much transcriber dependency to be able to compare results from different years.

We have generated a forced alignment of the hand-transcribed spoken text with the individual IHM data. This has been done at ICSI using ICSI-SRI speech-to-text system presented for the RT05s evaluation ([12]). The output of the same systems as in the previous table are evaluated using these new references and the results are shown in table 2

Evaluation	SDM		MDM	
campaign	non ovl.	ovl.	non ovl.	ovl.
RT02s	17.07%	19.07%	19.93%	21.89%
RT04s	12.84%	16.70%	13.98%	17.01%
RT05s	16.75%	19.34%	12.52%	15.06%
RT06s	23.86%	27.99%	16.46%	21.19%

Table 2: System DER on different evaluation campaigns using forced-alignment references

In general the missed-speech rate (indicative of multiple speakers where one is missed, and of missed speech labelled as non-speech) is much lower when evaluating with forced alignments. The difference between overlap and non-overlap results is much smaller and consistent using the forced alignments also, ranging between 2% and 4% of the total time.

In both RT02s and RT04s, the SDM condition outperforms the MDM condition in almost all cases. As pointed out, the MDM condition uses all available microphones in the room and obtains an enhanced signal by applying a delay-and-sum algorithm to them. Such a technique assumes that all microphones are of similar characteristics and therefore the MDM signal quality can only improve over that of any individual microphone. In some cases this might not be so, as one or more microphones might be of much lower quality than the others, thus degrading the signal. In table 3 we show the DER (including overlap speech) for the meetings in RT02s, RT04s and RT05s grouped by their origin. As we can see, the AMI meetings (containing 16 microphones) obtain a large gain in diarization by using all available channels. On the other hand, performance on the LDC meetings deteriorates significantly in some meetings when using all the microphones for MDM. The LDC meetings are mostly what causes the RT02s and RT04s results to show the SDM results to be better than MDM. All other meeting sources show an improvement on MDM versus SDM.

Source	# meetings	Average MDM	Average SDM
AMI	2	**3.78**%	12.66%
NIST	6	17.00%	19.98%
LDC	4	**30.22**%	18.75%
CMU	6	15.40%	16.83%
VT	2	12.27%	19.66%

Table 3: Average DER for MDM and SDM on a source basis

8. Conclusions

This paper presents the ICSI main submission to the RT06s speaker diarization evaluation campaign for the conference room task. This years' system contains four major improvements in the system to last year. They are: a new training-free speech/non-speech detector, a new initialization algorithm, an improved comparison between clusters previously purifying out non-discriminant frames, and the use of inter-channel delays as a feature in the diarization process. We show and analyze the results of this year's system on all meetings data available from previous and current evaluations and compare the results of the multiple distant microphones (MDM) and single distant microphone (SDM) conditions.

9. Acknowledgements

We would like to acknowledge the Speaker Diarization group at ICSI for their thoughtful comments and hard work, and Joe Frankel, Adam Janin for their help. This work was done during Xavier Angueras stay at ICSI within the Spanish visitors program.

10. References

[1] NIST rich transcription evaluations, website: http://www.nist.gov/speech/tests/rt.

[2] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *ICASSP'05*, Philadelphia, PA, March 2005, pp. 953–956.

[3] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *RT05s Meetings Recognition Evaluation*, Edinburgh, Great Brittain, July 2005.

[4] X. Anguera, C. Wooters, and J. Hernando, "Automatic cluster complexity and quantity selection: Towards robust speaker diarization," in *Speaker Odyssey 06*, Puerto Rico, USA (to appear), June 2006.

[5] ——, "Speaker diarization for multi-party meetings using acoustic fusion," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Puerto Rico, USA, November 2005.

[6] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando, "Hybrid speech/non-speech detector applied to speaker diarization of meetings," in *Speaker Odyssey 06*, Puerto Rico, USA (to appear), June 2006.

[7] X. Anguera, C. Wooters, and J. Hernando, "Frame purification for cluster comparison in speaker diarization," in *MMUA'06*, Toulouse, France (to appear), May 2006.

[8] S. Shaobing Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998.

[9] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *ASRU'03*, US Virgin Islands, USA, Dec. 2003.

[10] N. Mirghafori and C. Wooters, "Nuts and flakes: A study of data characteristics in speaker diarization," in *ICASSP'06*, Toulouse, France (to appear), May 2006.

[11] X. Anguera, C. Wooters, and J. Hernando, "Friends and enemies: A novel initialization for speaker diarization," in *ICSLP'06*, Pittsburgh, USA (submitted), September 2006.

[12] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The icsi-sri spring 2005 speech-to-text evaluation system," in *RT05s Meetings Recognition Evaluation*, Edinburgh, Great Brittain, July 2005.