Speaker Diarization for Multiple Distant Microphone Meetings: Mixing Acoustic Features And Inter-Channel Time Differences

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

[1] International Computer Science Institute, Berkeley CA 94708 USA,
[2] Universidad Politécnica de Madrid, 28040 Madrid, Spain
[3]Technical University of Catalonia, Barcelona, Spain
jpardo@die.upm.es,{xanguera,wooters}@icsi.berkeley.edu

Abstract

Speaker diarization for recordings made in meetings consists of identifying the number of participants in each meeting and creating a list of speech time intervals for each participant. In recently published work [7] we presented some experiments using only TDOA values (Time Delay Of Arrival for different channels) applied to this task. We demonstrated that information in those values can be used to segment the speakers. In this paper we have developed a method to mix the TDOA values with the acoustic values by calculating a combined log-likelihood between both sets of vectors. Using this method we have been able to reduce the DER by 16.34% (relative) for the NIST RT05s set (scored without overlap and manually transcribed references) the DER for our devel06s set (scored with overlap and force-aligned references) by 21% (relative) and the DER for the NIST RT06s (scored with overlap and manually transcribed references) by 15% (relative) .

Index terms: Speaker diarization, speaker segmentation, meetings recognition.

1. Introduction

There has been extensive research at ICSI in the last few years in the area of speaker segmentation and diarization [1],[2],[3],[4].

Speaker diarization for meetings consists of identifying the number of participants in each meeting and creating a list of speech time intervals for each participant. Notice that it may occur that two or more speakers talk at the same time, these overlap regions should be labelled with both speaker labels. It is important to emphasize that speaker diarization is done without using any knowledge about the number of speakers in the room, their location, the position and quality of the microphones, or the details of the acoustics of the room. These conditions make the task itself very tricky and very dependent on the characteristics of the room, the number of speakers and the number of channels. We have tried to make a system as robust as possible so its results are stable across different recording settings.

Speaker diarization for meetings using multiple distant microphones (MDM) should be easier compared to the use of a single distant microphone (SDM) for several reasons: a) there are redundant signals (one for each channel) that can be used to enhance the processed signal, even if some of the channels have a very poor signal to noise ratio; and b) there is information encoded in the signals about the spatial position of the source (speaker) that is different for each speaker.

In previous work [9], a processing technique using the time delay of arrival (TDOA) was applied to the different microphone channels by delaying in time and summing the channels to create an enhanced signal. With this enhanced signal, the speaker diarization error (DER) was improved by 3.3% relative compared to the single channel error for the RT05s evaluation set, 23% relative for the RT04s development set, and 2.3% relative for the RT04s evaluation set (see [10] for more information about the databases and the task).

While in the work mentioned above, improvements were obtained, no direct information about the delays between different microphones was used in the segmentation and clustering process.

In recent work [7], we processed the TDOA values and clustered them to obtain a segmentation hypothesis. Using only this information we obtained a 31.2% diarization error rate (DER) for the NIST´s RT05s conference room evaluation set. For a subset of NIST´s RT04s, we obtained 35.73% DER error (not including False Alarms in this case). Comparing those results with the ones presented by Ellis and Liu [8], who also used inter-channel differences for the same data, we obtained 43% relative improvement..

In this paper we combine the acoustic front end features (MFCC) with the TDOA features to obtain an enhanced segmentation useful for this task. Including the TDOA values we have been able to improve baseline results by 16.35% relative for the RT05s evaluation set 21% relative for the devel06s database (see the explanation of the database content below) and 15% relative for the RT06s evaluation set.

2. System Description

The basic procedure is based on the segmentation and clustering proposed in [2],[3] using only acoustic features, without the use of the purification method mentioned in [3]. But there are substantial differences as explained below.

2.1. Speech/non speech calculation

As a first step in the diarization process non-speech frames are identified and removed. We have used the SRI speech/non-speech detector [4] or a more recent system developed at ICSI [12].

2.2. Delay generation

We calculate the cross-correlation between the signals coming from the different channels and estimate the TDOA as the maximum of the cross-correlation function. The details of the delay generation procedure are described in [9].

For a set of microphones, we choose the microphone with overall best cross-correlation with all others as the reference microphone and calculate the delay of the signals coming to the other microphones relative to the reference microphone. We form a vector of these delays that has as many components as the number of microphones minus 1. We use a window width of 500 msec with a shift of 10 msec.

2.3. Acoustic feature extraction

The signals coming from the different microphones are delayed and added together to form a single enhanced signal[9]. On the enhanced signal we calculate a vector of 19 MFCC coefficients using a 30 msec analysis window and a frame shift of 10 msecs.

2.4. Initialization

The initialization requires a "guess" at the maximum number of speakers (K) that are likely to occur in the data. The data are then divided into K equal-length segments, and each segment is assigned to one model. Each model's parameters are then trained using their assigned data. With the trained models we segment the data (using the Viterbi algorithm) and retrain them over several iterations. The clustering process uses of an ergodic HMM model that has a number of states equal to the initial number of clusters (K). Each state in the HMM contains a sequence of MD subtates which are used to impose a minimum duration. Within a state, each one of the sub-states shares a probability density function (PDF) modelled with a Gaussian mixture model (GMM) with a diagonal covariance matrix. Each GMM starts with "g" gaussians which are changed later in the merging process. The models for the acoustic vectors and for the delay vectors are trained in parallel but kept as separate models. The number of initial gaussians per model is different for the acoustic vectors and for the delay vectors. In previous work [7] we did some experimentation using only the delay vectors in the segmentation and clustering procedure. Using 10 initial clusters each starting with a single we obtained the best results for the MDM RT05s set (31.2% DER). For the acoustic features we use 5 initial gaussians per model.

The combined log-likelihood Clog for each state and every frame is obtained by combining the log likelihoods from the acoustic vectors and the log likelihoods from the delay vectors using the following formula:

$$C \log\ p(x[i], y[i] \mid \theta_a) =$$
$$\alpha \log\ p(x[i] \mid \theta_{ax}) + (1-\alpha) \log\ p(y[i] \mid \theta_{ay})$$

(Eq.1)

θ_a is the compound model for cluster a, θ_{ax} is the model created for cluster *a* using the acoustic vectors *x[n]* and θ_{ay} is the model created for cluster a using the delay vectors *y[n]*. α is a weight that has to be determined by some method. Currently we determine it empirically using development data.

2.4.1. Clustering process

The initialized modelseed the clustering and segmentation processes described next.

The iterative segmentation and merging process consists of the following steps:

1. Run a Viterbi decode to re-segment the date.
2. Retrain the models using the segmentation from (1).
3. Select the pair of clusters with the largest merge score (Eq. 2)> 0.0 (Since Eq. 2 produces positive scores for models that are similar, and negative scores for models that are different, a natural threshold for the system is 0.0)
4. If no pair of clusters is found, stop.
5. Merge the pair of clusters found in (3). The models for the individual clusters in the pair are replaced by a single, combined model.
6. Go to (1).

2.4.2. Merging score

One of the main problems in the segmentation and clustering process is deciding which merging score to use. The BIC criterion has been used extensively, giving good results [1],[11] and the modification of BIC to eliminate the need of a penalty term has also given us good results. Nevertheless it is still an open question as to how much the performance depends on the kind of data vectors and models that are used in the comparisons. The modified BIC that we use for merging is:

$$\Delta BIC = C \log\ p(D \mid \theta) -$$
$$C \log\ p(D_a \mid \theta_a) - C \log\ p(D_b \mid \theta_b)$$

(Eq.2)

θ_a is the model created with D_a and θ_b is the model created with D_b, θ is the model created with D which is the union of D_a and D_b , the key to this modified BIC is that the number of parameters in θ must equal the sum of the number of parameters in θ_a and θ_b.

3. Experiments and Results

3.1. Experiments with RT05s set and hand labeled references (System A).

We have used the RT05s MDM conference meetings evaluation data in our experiments. The data consists of 10 meetings from which 10 minute excerpts have been extracted [10]. The DER was obtained using the standard NIST procedure comparing the segmentation results with the hand labelled reference data. In the first column of Table 1, results for the independent systems and for the combined system are shown. We have not included overlapping speech in the score calculation. For the results presented here we have used the SRI speech/non-speech detector.

The results obtained by using the combined system give a relative error improvement of 16.34 % DER.

We have used a weight factor α=0.9. The first question to answer is how to determine α. Fig 1 shows a plot of DER as a function of α. It can be seen that a badly chosen weight factor

can seriously degrade performance, since the delays alone have a much worse performance than the acoustic vectors alone.

Features used	DER eval05s (system A)	DER devel06s (system B)
Delays only	31.20 %	31.97 %
Acoustic features only	18.48 %	12.71 %
Combined acoustic+delays	15.46 %	10.04 %
Relative error reduction	16.34 %	21 %

Table 1: DER error for the eval05s and devel06s dataset obtained using acoustic features only, delay features only and combined features using System A and System B

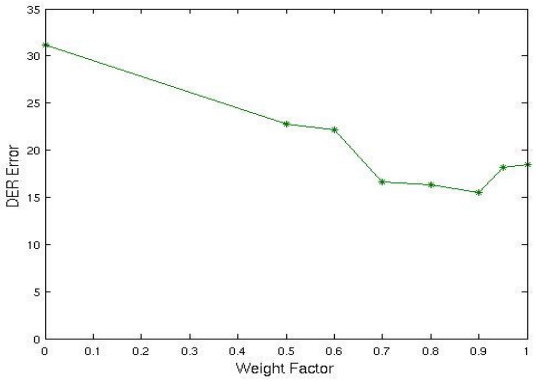

Figure 1: *Plot of different DER errors as a function of the weight factor applied for the Eval05s set*

3.2. Experiments with force-aligned references (System B)

For the NIST RT06s evaluation campaign we decided to select a set of development shows from all previous data sets: RT02s, RT04s, and RT05s. The set of shows are given in Table 2. We we will refer to this set of shows as the devel06s data. For the RT06 evaluation we have made several changes compared to what has been mentioned above:

- Since the evaluation this year was going to be made taking into account the overlap between speakers, all the new scoring has been done taking overlaps into account[1].
- We have discovered that the hand-aligned data contained a lot of non-speech events (breaths, cough, lipsmacks etc), especially at the beginning and end of every speaker turn. Also and sometimes overlap was marked when there was. For this reason we decide to use references obtained by using the SRI recognizer to force-align the data.
- We have used a new speech/non-speech detector that doesn´t need training data[12].

The results on the devel06s are presented in Table 2. In the different columns we present the percentage of missed speech, false alarm speech, speaker error and total diarization error. There is a missed if one (or several) speakers talking at the same time are not labeled. There is a false alarm if the system assigns a label to a region where there is no speech. There is a speaker error if the label assigned by the system does not match the tgrue speaker (see [10]).

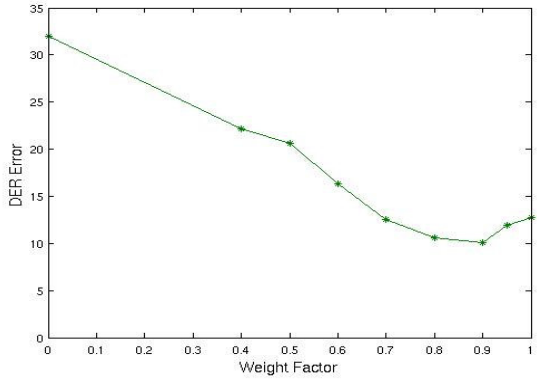

Figure 2*: DER as a function of the weight factor for the devel06s data)*

File	Miss	FA	Spkr	Total
AMI_20041210-1052	0.40	1.20	1.10	2.69
AMI_20050204-1206	2.60	2.20	3.30	8.01
CMU_20050228-1615	9.30	1.20	1.80	12.30
CMU_20050301-1415	3.70	1.60	1.10	6.41
ICSI_20000807-1000	4.60	0.40	3.80	8.77
ICSI_20010208-1430	3.60	1.10	11.00	15.72
LDC_20011116-1400	2.10	3.00	4.20	9.32
LDC_20011116-1500	5.90	1.10	7.60	14.65
NIST_20030623-1409	1.00	0.70	1.40	3.08
NIST_20030925-1517	7.70	5.70	9.60	22.95
VT_20050304-1300	0.60	1.00	2.80	4.43
VT_20050318-1430	1.30	6.20	13.80	21.36
ALL	**3.40**	**1.90**	**4.70**	**10.04**

Table 2*: Results for the subset of shows listed. We present the percentage of Missed speech, False Alarm speech,, Speaker error and total Diarization error (DER).*

The results presented for devel06s in Table 1 and Table 2, used K=16 and a minimum duration of 2.5 seconds. Again it can be seen that performance compared to the baseline system (without

[1]Scoring with overlap means taking into account the regions where more than one speaker talks at the same time and an error is counted if any of the speakers is not found.

using the delays information) is improved by using the delays information. In the second column of Table 1, the DER using acoustic features, delay features and the combination of both are presented. The relative improvement of the combined system compared to the acoustic features alone is 21 %.

In Figure 2, we show the DER as a function of the weight factor applied. Again a good weight factor is crucial to obtaining good results.

3.3. Official results at NIST RT06

This system[1] was presented as a contrastive system in the official NIST RT 06s evaluation campaign giving a DER of 35.77%. Although the original plan was to score the data using force-aligned labels, the results were finally scored using hand-made references. The scoring with force-aligned labels gave a DER of 20.03%. We calculated the after-eval score using only acoustic features and hand-made references and it gave us a DER of 42.13%. Thus the use of delay information reduced the error of the system by 15% relative.

4. Discussion

In all the experiments that we have done using this method we have been able to improve the results obtained using only the acoustic vectors. The results on both RT05s set, devel06s set and RT06s set show substantial improvements.

The problem of integrating TDOA information with acoustic information is not trivial. Previous experiments merging both types of information in a single long vector yielded poorer performance. We believe that this may be due to the use of diagonal covariance matrices in a non-homogeneuos vector.

There is a large amount of work ahead researching methods for merging both sources of information, especially since we do not know yet if the merging metric used is the best possible metric. There is also a need to develop methods to estimate the weight factor automatically.

5. Conclusions

In this paper we have presented a method to combine acoustic features and delay features to improve speaker diarization performance. The results are significantly better than the ones obtained using acoustic features alone. There are still many unknowns to the method (some of them inherent to the clustering procedure) such as how to choose the minimum duration constraint, how to choose the initial number of clusters, and how to choose the initial number of Gaussians for each cluster. Particularly important is how to select a good weight factor between the acoustic features and the delay features that is robust and generalizes to different conditions of rooms, number of speakers, number of microphones etc.

6. Acknowledgements

This work was supported by the Joint Spain-ICSI Visitor Program and by the projects ROBINT (DPI 2004-07908-C02) , TINA (UPM-CAM R05-10922) and EDECAN (TIN2005-08660-C04). We also would like to thank Andreas Stolcke, Kemal Sönmez and Nikki Mirghafori for many helpful discussions. We appreciate the help of Michael Ellsworth in reviewing the English.

7. References

[1] J. Ferreiros, D. Ellis: Using Acoustic Condition Clustering To Improve Acoustic Change Detection On Broadcast News. Proc. ICSLP 2000

[2] J. Ajmera, C,Wooters : A Robust speaker clustering algorithm, IEEE ASRU 2003.

[3] X. Anguera, C. Wooters, B. Pesking and Mateu Aguiló : Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System, Proc NIST MLMI Meeting Recognition Workshop, Edinburgh, 2005

[4] C.Wooters, J, Fung, B. Pesking, X. Anguera, "Towards Robust Speaker Segmentation: The ICSI-SRI Fall 2004 Diarization System" NIST RT-04F Workshop, Nov. 2004.

[5] 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" Proceedings of NIST MLMI Meeting Recognition Workshop, Edinburgh.

[6] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters and B. Wrede, "The ICSI Meeting Project: Resources and Research" NIST ICASSP 2004 Meeting Recognition Workshop, Montreal

[7] J.M.Pardo, X. Anguera, C. Wooters: Speaker Diarization For Multi-Microphone Meetings Using only Between-Channel Differences. Proc. MLMI 06, 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 1-3 May 2006, Washington DC, USA. To appear in Lecture Notes in Computer Science.

[8] D.P.W Elis and Jerry C.Liu : Speaker Turn Segmentation Based On Between-Channels Differences, Proc. ICASSP 2004.

[9] X. Anguera, C. Wooters, J. Hernando : Speaker Diarization For Multi-Party Meetings Using Acoustic Fusion, IEEE ASRU, 2005.

[10] NIST Spring 2005 (RT05S) Rich Transcription Meeting Recognition Evaluation Plan, http://www.itl.nist.gov/iad/894.01/tests/rt/rt2005/spring/

[11] S.S. Chen, P.S. Gopalakrishnan: Speaker Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion, Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Virginia, USA, Feb. 1998.

[12] X. Anguera, M Aguiló, C. Wooters, C. Nadeu, J. Hernando "Hybrid Speech/non-speech detector applied to Speaker Diarization of Meetings" IEEE Odyssey 2006: The Speaker and Language Recognition Workshop 28-30 June 2006 San Juan, Puerto Rico

[1] With a slightly modified delay calculation method