Speaker Diarization for Multi-microphone Meetings Using Only Between-Channel Differences

Jose M. Pardo[1,2], Xavier Anguera[1,3], and 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,xanguera,wooters}@icsi.berkeley.edu

Abstract. We present a method to extract speaker turn segmentation from multiple distant microphones (MDM) using only delay values found via a cross-correlation between the available channels. The method is robust against the number of speakers (which is unknown to the system), the number of channels, and the acoustics of the room. The delays between channels are processed and clustered to obtain a segmentation hypothesis. We have obtained a 31.2% diarization error rate (DER) for the NIST´s RT05s MDM conference room evaluation set. For a MDM subset of NIST´s RT04s development set, we have obtained 36.93% DER and 35.73% DER*. Comparing those results with the ones presented by Ellis and Liu [8], who also used between-channels differences for the same data, we have obtained 43% relative improvement in the error rate.

1 Introduction

There has been extensive research at ICSI in the last few years in the area of speaker segmentation and diarization [1,2,3,4,5,6,7]. Speaker diarization is the task of identifying the number of participants in a meeting and create a list of speech time intervals for each such participant.

The task of speaker diarization for meetings with many speakers and multiple distant microphones (MDM) should be easier compared to the use of a single distant-microphone (SDM) because: 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 from one to another. 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).

It is important to emphasize that the task is done without using any knowledge about the number of speakers in the room, their location, the locations and quality of the microphones, or the details of the acoustics of the room.

S. Renals, S. Bengio, and J. Fiscus (Eds.): MLMI 2006, LNCS 4299, pp. 257 – 264, 2006.

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 order to study and analyze the information contained in the delays, we have performed some experiments to determine to what extent the delays by themselves can be used to segment and cluster the different speakers in a room. We have tried to develop a system that is robust to the changes in the meeting conditions, room, microphones, speakers, etc.

The only work of which we are aware that only uses between-channel differences for speaker turn segmentation is the work of Ellis and Liu [8]. In their work, they used the cross correlation between channels to find a peak that represents a delay value between two channels. They later clustered the delay values to create segments in the speech frames. The results they reported for the set of shows corresponding to the RT04s development set is 62.3% DER* error.[1] We present a method to use only the delays to obtain a segmentation hypothesis. Using our method, we obtain a diarization error (DER*) [10] of 35.73% for the same set of shows. Furthermore, for the set of shows corresponding to the RT05s MDM conference evaluation set, we have obtained a 31.2% DER error. The DER error could be reduced further, since one of the shows had a large number of false alarm speech errors (due to big background noises such as papers rustling, etc). Without taking this show into account, the average DER error rate for the RT05s set goes down to 27.85%.

The paper is organized as follows: In Section 2 we describe the basics of our system, in Section 3 we describe the experiments done, in Section 4 we discuss the results, Section 5 finishes with our conclusions.

2 Description of the System

2.1 Delay Generation

Given any two microphones (i and j) and one source of speech (x[n]), let us call the signals received by each microphone $x_i[n]$and $x_j[n]$.

We define the delay of $x_i[n]$ with respect to $x_j[n]$ as the time difference of the sound arriving at each microphone.

If we assume the produced wave-front is flat when reaching the microphones, and further assume a non-dispersive wave propagation, we obtain the delay (in # of samples) as

$$d(i,j) = \frac{D(i,j)\cos\alpha}{c.f_s} \tag{1}$$

Where D(i,j) is the distance between the two microphones, α is the angle of arrival of the source speech, c is the speed of sound (in m/sec) and f_s is the sampling frequency (in samples/sec.), see [9].

In order to estimate the TDOA between segments, we cannot directly use equation (1) because we do not know the number of speakers nor their locations. We used a

[1] The equivalent that they used is DER minus False Alarm (in NIST terminology), we called it DER*.

modified version of the Generalized Cross Correlation with phase transform ($GCC_{PHAT}(f)$) (see [11]) and estimate the delays between microphones with the following formula:

$$d(i,j)=\arg\max_{d}(R_{PHAT}(d)) \tag{2}$$

$R_{PHAT}(d)$ is the inverse transform of $G_{PHAT}(f)$ (the generalized cross correlation)

For a set of microphones, we choose any microphone 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 per frame. Non-speech frames are estimated with the SRI Meetings speech/non-speech detector and are excluded from the subsequent process see [5]. All the data given below about speech/non-speech errors exclusively originate from this system.

2.2 Segmentation and Agglomerative Clustering

The segmentation and clustering is very similar to what is proposed in [3] for segmentation and clustering using acoustic features. We use the vectors explained above to feed the initial segmentation and posterior resegmentation and clustering as proposed in [3]. Essentially, the process consists of two modules: the initialization and the clustering. 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 its assigned data. To model each cluster, we use and HMM consisting of a minimum number of states all with the same output pdf -a gaussian mixture- with a diagonal covariance matrix starting with "g" gaussians per model. These are the models that seed the clustering and segmentation processes described next.

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 [1,12] and the modification of BIC to eliminate the need of a penalty term that compensates for different number of parameters has given us also good results 3], although it is an important open question how much it depends on the kind of data vectors and models that are used in the comparisons.

The modified BIC is the following:

$$\log p(D/\theta) \geq \log p(D_a/\theta_a) + \log p(D_b/\theta_b) \tag{3}$$

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, with the number of parameters in θ equal to the sum of the number of parameters in θ_a plus the number of parameters in θ_b

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

1. Run a Viterbi decode to re-segment the data.
2. Retrain the models using the segmentation from (1).
3. Select the pair of clusters with the largest merge score (Eq. 3)> 0.0 (Since Eq. 3 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).

3 Experiments and Evaluation

We have used the RT05s MDM conference meetings evaluation data in our initial development experiments. The data consists of 10 meetings from which 10 minutes excerpts for every one have been extracted [10]. Several combinations of the parameters "g" and "K" have been tried, with the best results obtained using 1 mixture and 10 initial clusters. The speaker diarization errors obtained with several combinations of these parameters are presented in Table 1.

Table 1. Speaker diarization errors DER for the RT05s MDM conference room eval set

	# of initial clusters	
# of gaussians	**10**	***20***
1	31.20 %	34.77 %
2	38.68%	43.49%

The breakdown of these data (1 gaussian, 10 initial clusters) into different shows is presented in Table 2. We show the Missed Speech error, the False Alarm Speech error, the Speech/NonSpeech error (SpNsp), the Speaker error and the overall DER error[2]. In the results presented, the regions where more than one speaker are talking have been excluded[3]. We have analyzed the results and found that the show VT_20050318-1430 has a big SpNsp error, and particularly the False Alarm error. This is due to a background paper noise that is erroneously detected as speech by the SRI system. Without taking into account this show, the average DER is 27.85%. We have also investigated the minimum error (ORACLE) that could be obtained by this procedure by using the clustering iteration loop without any stopping criterion and calculating the theoretical error obtained if the system stopped after each iteration.

[2] The speech/non-speech segmentation is not calculated by our system and it is presented here for completeness.
[3] This condition is considered in the evaluation tool as the "no overlap" condition.

The DER (ORACLE) error for these shows (not including VT_20050318-1430) is 23.28%. This error is just a way of measuring the possible absolute limit for our current experiments if an optimum stopping criterion were known.

Table 2. Missed speech, False Alarm speech, Speech/Non-speech error and Diarization error for the RT05s eval set using 1 mixture and 10 initial clusters

File	Miss	FA	SpNsp	Spkr	Total
AMI_20041210-1052	1,1	1,9	3	13,5	16,53
AMI_20050204-1206	1,8	1,7	3,5	19,6	23,03
CMU_20050228-1615	0,1	1	1,1	17,2	18,28
CMU_20050301-1415	0,2	3,3	3,5	42,4	45,88
ICSI_20010531-1030	4,3	1,3	5,6	15	20,59
ICSI_20011113-1100	2,9	2,7	5,6	39,9	45,52
NIST_20050412-1303	0,6	2,9	3,5	21,7	25,19
NIST_20050427-0939	1,5	2,5	4	33,2	37,18
VT_20050304-1300	0	3,6	3,6	22,1	25,7
VT_20050318-1430	0,3	22,6	22,9	38,4	61,27
ALL	1,3	4	5,3	25,9	31,2

For the purpose of comparison of this method with the regular method, we have run the same standard procedure but now using the normal MFCC feature set and we obtained a DER error of 13.38% for the same set of shows (not including VT_20050318-1430 for the reasons mentioned above)[4]. This data shows us that there is still a big gap between the errors obtained using only time differences and the ones obtained using only acoustic data.

Table 3. Comparisons between results obtained by Ellis and Liu and our results in the same subset of shows from NIST RT04 development data

Meeting	Ellis DER*	Our System DER*	Our System DER	Number of microphones used
LDC_20011116-1400	66%	6.89%	8.89%	4
LDC_20011116-1500	77.3%	59.33%	59.63%	4
NIST_20020214-1148	58%	33.32%	37.72%	4
NIST_20020305-1007	46.1%	32.81%	34.11%	4
ICSI_20010208-1430	49.1%	29.9%	38.7%	4
ICSI_20010322-1450	63.3%	43.53%	43.83%	4
Average All	**62.3%**	**35.73%**	**36.93**%	

[4] It should be noted however that these results have been obtained using the Standard system presented at the NIST RT05s meeting BUT without the purification system included[3].

To further demonstrate that there is information in the timing differences between channels, we ran an experiment using just random numbers and processed them to extract the diarization error. For the show processed in this manner, ICSI_20010531-1030 we obtained 93.23% DER error compared to 21.14% DER error using the same parameters settings. The system is able to find information in the time differences between signals coming from different microphones.

In order to be able to compare our results with the ones presented by Ellis and Liu [8], we have run the system with the same set of shows that they used in their experiments, and reducing the number of channels to 4 in all cases. In Table 3, the comparisons of both experiments are presented. It is important to notice that in these results, two of the shows from NIST RT04 (the CMU shows) have not been used because the conditions of these shows (only one distant microphone) are not compatible with the conditions of our experiment (multiple distant microphones)[5]. Also the results presented here include the overlap regions and no False Alarms (we call it the DER* error). We have included in Table 3 also the standard DER error for completeness. The analysis of the results show a big improvement of our system compared to the Ellis one. The differences may well come from the different way we use to calculate the delays between signals and the different segmentation and clustering procedure. Since the number of microphones used in this experiment were less than the number of microphones available, we have also investigated the error rate that we could obtain for the same set of shows if we used all the available microphones. Table 4 gives results of this comparison. It can be seen that the use of more microphones reduces the DER error rate by 3.26% absolute.

Table 4. Comparisons between DER rates obtained using 4 channels and results using all the channels available in the system

Meeting	# microphones used	Diarization error	# microphones used	Diarization error
LDC_20011116-1400	4	8.89%	7	12.26%
LDC_20011116-1500	4	59.63%	8	45.72%
NIST_20020214-1148	4	37.72%	7	36.40%
NIST_20020305-1007	4	34.11%	6	41.37%
ICSI_20010208-1430	4	38.7%	6	19.81%
ICSI_20010322-1450	4	43.83%	6	44.68%
Average All		**36.93%**		**33.67%**

4 Discussion

The estimation of errors in the Ellis system was performed with a quantized version of the scoring method that we have used (the official NIST scoring program). In his

[5] Ellis and Liu developed an artificial condition for those two shows that do not make sense in our method. Those two shows are then not used.

scoring the errors were quantized in segments of length 250 msec and no reference was made to the forgiveness collar of 250 msec at each side of a reference segment that was done in the NIST scoring software. Also, from the explanations given in their paper, they did not count the regions of silence in the reference transcriptions. We have discounted those errors in our data and defined DER* (see column 3 of Table 3).

If we compare our results to Ellis results, there is an important improvement. Our experiments further support the Ellis and Liu idea that there is information in the timing differences between different channels that can be used to extract speaker turn information (obvious in any case but usually difficult to extract). However if we compare the results that we obtain with the results obtained with our standard spectral system there is still a big gap to cover. Nonetheless, in this paper we just wanted to show that there is information in the timing differences between channels that could be used in speaker diarization systems. It is our purpose to continue research in this area in order to be able to integrate information coming from different sources and apply it to this task.

5 Conclusions

In this paper we have presented some experiments to analyze the information that exists in the timing differences between channels in the speaker diarization task for multiple distant microphones. While our results are significantly better than the ones published up to now with the same type of information, these results should be considered as a first step towards the development of improved systems for speaker diarization in the presence of multiple microphones.

Acknowledgements

This work was supported by the Joint Spain-ICSI Visitor Program. We also would like to thank Andreas Stolcke, Kemal Sönmez and Nikki Mirghafori for many helpful discussions. We acknowledge the help of Adam Janin in reviewing the paper.

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, N. Mirghafori, A. Stolcke, T. Pirinen, I Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin and M. Ostendorf :"The 2004 ICSI-SRI-UW Meeting Recognition System" In Proceedings of the Joint AMI/Pascal/IM2/M4 Workshop on Meeting Recognition. Also published in Lecture Notes in Computer Science, Volume 3361 / 2005.
5. 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.

6. 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.
7. 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
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. M.S. Brandstein and H.F. Silverman: A Robust Method For Speech Signal Time-Delay Estimation In Reverberant Rooms, ICASSP 97, Munich
12. 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