Speaker Diarization for Meetings Using Acoustic Fusion

Xavier Anguera[1,2], Chuck Wooters[1], Javier Hernando[2]

[1] International Computer Science Institute
1947 Center st., Suite 600
Berkeley, CA 94704, U.S.A.
[2] Technical University of Catalonia (UPC)
Jordi Girona 1-3, building D5
08034 Barcelona, Spain
{xanguera,wooters}@icsi.berkeley.edu

Abstract

This paper proposes the use of delay-and-sum beamforming to improve speaker segmentation on meeting data using multiple distant microphones (MDM). The time delay of arrival (TDOA) between each channel and a reference is computed using a small scrolling window. No information about the location or setup of the microphones used to make the recordings is needed. A postprocessing of the delays is used in an attempt to ensure their continuity given silence, noises and overlapped speech. The resulting signal is then clustered using ICSI's Broadcast News Diarization system. Tests on the RT04s Meetings evaluation database show competitive performance with respect to published results on the same data. Additionally, the proposed implementation running on the MDM task shows a relative improvement of almost 30% over the single distant microphone (SDM) task.

1. Introduction

Speaker diarization attempts to answer the question "Who spoke when?" in a multi-person recording. Most of the research addresses the blind situation, where no information about the number of speakers or their identities is known.

In recent years there has been extensive research in speaker diarization for the Broadcast News (BN) environment ([1]) where single channel recordings from radio and TV include speech in various acoustic environments, music, and advertisements. Recent research is being conducted in the Meetings environment (due to the ongoing projects AMI, CHIL and IM2) where speaker diarization encounters many differences. One difference between the BN domain and the Meetings domain is that for meetings, more than one microphone may be available for processing. These microphones are typically located in the middle of a meeting table and are of lower quality than the microphones used in BN. Processing data from these microphones is referred to as the Multiple Distant Microphones (MDM) task. Other differences between speech from meetings and speech from BN include: speech in meetings is spontaneous, there are more silence segments, and often, there is more than one speaker talking at the same time.

Due to the novelty of the task, very few publications have addressed the problem. The baseline approach is to consider only the best channel (usually the most centrally located) and perform speaker diarization on it. This is done in [2] for the RT04s evaluation. In order to use the information from all channels in [3] a Speech Activity Detector (SAD) is used to split the channels into speaker homogeneous segments and a single reconstructed channel is created by selecting the best segment at each instant (according to SNR and Energy). Diarization is then done on this reconstructed channel. This system doesn't address the problem of overlapping speech that results in more than one speaker per segment, and ultimately only one channel's data is being used for the diarization, ignoring any information from the rest. Another option is to independently cluster all the channels and then postprocess the resulting segmentation. In [4] an iterative process is used looking for the longest speaker intervention from all channels.

In this paper we present an approach from the point of view of signal processing, where a classic delay-and-sum (D&S) is used to combine all channels into one single enhanced channel, that is then clustered using the ICSI-SRI Broadcast News Speaker Diarization system ([5]). By using the D&S algorithm incrementatlly, we attempt to align all the channels with respect to the speaker who is currently talking. This has the effect of improving the SNR of the resulting channel with respect to the individual channels, regardless of the location of the speaker. To perform D&S we use a scrolling window through the signal and compute the Time Delay of Arrival (TDOA) of each window using GCC-PHAT ([6]). Three different filtering techniques are employed to smooth the computed TDOA to avoid instabilities due to overlapped speech, silence segments, or degraded channels.

In Section 2, the D&S and TDOA estimation theory is revisited. In Section 3 we present the system implementation. Finally, in Section 4 we discuss experiments and results.

2. Delay-and-Sum in Meetings

The delay-and-sum beamforming technique ([7]) is a simple yet effective way to enhance an input signal when the signal has been recorded on more than one microphone. It doesn't assume any information about the position of the microphones or their placement. The principle of operation of the D&S can be seen in Figure 1.

Given any two microphones (i and j) and one source of speech ($x[n]$), the signals received are $x_i[n]$ and $x_j[n]$. Considering only additive noise ($n_i[n]$ and $n_j[n]$) and one speaker talking, we have:

$$\begin{aligned} x_i[n] &= x[n] + n_i[n] \\ x_j[n] &= x[n - d(i,j)] + n_j[n] \end{aligned} \quad (1)$$

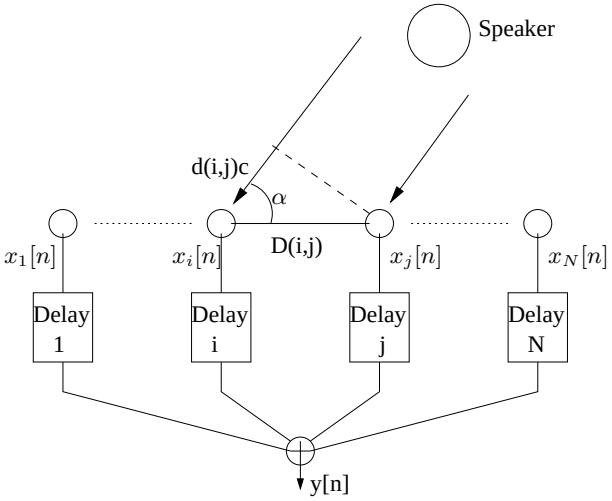

Figure 1: *Delay-and-sum system*

We define the delay of x_i with respect to x_j ($d(i,j)$) as the time difference of the sound arriving at each microphone. If we consider the produced wave front flat when reaching the microphones, and nondispersive wave propagation, we obtain the delay as

$$d(i,j) = \frac{D(i,j) \cdot cos\alpha}{c} \tag{2}$$

Where $D(i,j)$ is the distance between the two microphones, α is the angle of arrival of the source speech and c is the speed of sound.

Given N microphones, if we know their delay with respect to a reference microphone x_0, we can obtain an enhanced signal with

$$y(n) = x_0(n) + \sum_{i=1}^{N-1} x_i(n - d(0,i)) \tag{3}$$

By adding together the aligned signals the usable speech adds together and the ambient noise (if we consider it random) is reduced. Using the D&S we can obtain up to a 3db SNR improvement each time that we double the number of microphones.

2.1. TDOA Estimation via GCC-PHAT

In order to estimate the TDOA between two segments from two microphones we use a modified version of the Generalized Cross Correlation (GCC) called the generalized cross correlation with phase transform (GCC-PHAT) method (see [6]). Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is defined as:

$$G_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|} \tag{4}$$

Where $X_i(f)$ and $X_j(f)$ are the fourier transforms of the two signals and $[]^*$ denotes the complex conjugate. The TDOA for these two microphones is estimated as:

$$\hat{d}_{PHAT}(i,j) = \underset{d}{argmax} \left(\hat{R}_{PHAT}(d)\right) \tag{5}$$

Where $\hat{R}_{PHAT}(d)$ is the inverse fourier transform of $G_{PHAT}(f)$. Although the maximum value of $\hat{R}_{PHAT}(d)$ corresponds to the estimated TDOA, we have found it useful to keep the top N values for further processing.

3. System Implementation

We can see in figure 2 the basic blocks forming the system presented in this paper. The raw signal coming from the different available channels are individually Wiener-filtered to improve their SNR in the same way as was done in the ICSI-SRI-UW Meetings recognition system ([8]). Then the D&S is performed resulting in one output channel[1]. This channel is fed into the diarization system that outputs the desired segmentation of the meeting. The diarization system used is the same as presented in the Broadcast News task RT04f, described in [5].

For each meeting, regardless of the number of microphones available, the most centrally located microphone[2] (as defined by NIST in the RT04s evaluation for each meeting) is used as a reference channel.

Then a segment of 0.5s is obtained for each channel and the reference and they are weighted with a Hamming window prior to computing the TDOA. The selection of the analysis window is important and needs some consideration. If we choose a very small window we encounter problems when estimating the delay between channels. Additionally, smaller segments are more likely to contain mostly or only silence/noise, resulting in erroneous TDOAs. On the other hand, if the analysis window is too big we will miss short speaker turns, assigning them a wrong TDOA, and won't accurately detect the speaker change points. Experimentally, we have found a value of 500ms to be a good tradeoff for both cases and for system speed.

3.1. Robust TDOA Estimation

To obtain the TDOA for a given analysis window, the GCC-PHAT ([6]) is computed between the segments of each channel and the reference. Such a measure is more robust and accurate than cross-correlation when the noise level is low and it outputs values normalized from 0 to 1. The 8 major peak values are found in a region of allowed delays of $\pm$20ms (which corresponds to 7 meters of separation between microphones in the worse case scenario). As our aim is to enhance the output signal and not to obtain an optimum estimate of the position of each speaker through their TDOA, the following techniques are applied to the delays:

1. Confidence level: The TDOA estimation in silence regions or in low SNR regions is not reliable and often is completely wrong. So we threshold the max GCC-PHAT value using a cutoff of 0.1 (10% of the maximum value possible). When the max GCC-PHAT value doesn't exceed this threshold, we let d(n) = d(n-1).

[1]Note that this is similar to the technique that we used in our Rich Transcription Meetings system ([8]). The differences are in the window size used, and the way that the TDOA is obtained.

[2]We performed tests using the microphone with greater RMS but that resulted in poorer overall performance.

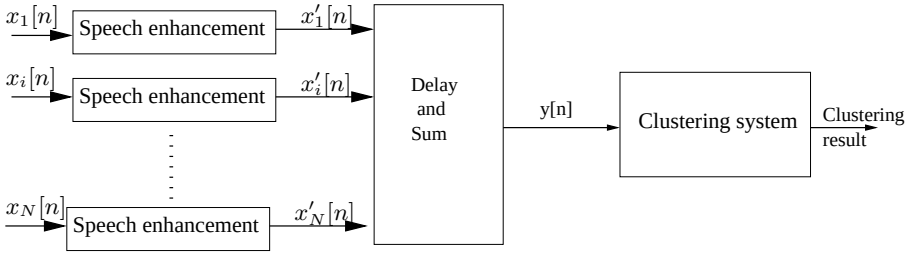

Figure 2: *Proposed diarization system for meetings data*

2. TDOA Continuity: In the presence of multiple speakers or impulsive noises it can occur that the main peak of the GCC-PHAT computation changes from one source to another between analysis windows. This filtering step tries to enforce continuity of the TDOA values by trying to find a $d(n) = d(n-1) \pm \Delta$ within all computed maxima, where $\Delta = 10$ samples in our implementation.

3. Correlation Continuity: When a new source is detected (non continuity of TDOA) a further step to the confidence level is to enforce a continuity of the GCC-PHAT correlation values given the previously detected TDOA's. It acts as a dynamic confidence thresholding dependent on the show. We do so by keeping the online causal mean of the GCC-PHAT correlation values and and checking that $corr(n) > \overline{corr}(n-1)$. otherwise we set d(n) = d(n-1).

The obtained TDOA estimates are applied to each windowed segment. The segments are then summed together with equal weights. In order to reconstruct the entire signal, an overlap of 50% is applied to the resulting segments, using a triangular windowing to obtain an overall constant gain.

3.2. Diarization System

The enhanced signal, created as described in the previous section, is "diarized" using the ICSI Broadcast News speaker diarization system. This system is a bottom-up agglomerative system that uses the Bayesian Information Criterion as a merging and stopping criteria. The algorithm is initialized with 10 clusters of equal size from the enhanced input signal and iteratively reassigns the acoustic observations and merges the clusters until obtaining the optimal number of clusters according to the BIC criterion. This system does not need any pre-trained models or threshold adjustments. Thus porting it to a new task is straightforward. For more information see [5].

4. Experiments and Results

Speaker diarization experiments were conducted using the databases distributed for the Rich Transcription 2004 Spring Meeting Recognition Evaluation, RT04s ([1]). For our experiments, we used the development and evaluation data, containing approximately 80 and 90 minutes of speech, respectively. The data consists of 10 and 11 minute excerpts from meetings collected at four different sites. In this task only the distant microphones (those mounted on the table tops during the meetings) are used for clustering. We experimented with two conditions-MDM (Multiple Distant Microphones), where all microphones can be used, and SDM (Single Distant Microphones), where only the most centrally located microphone is used. We use the SDM case as a baseline to compare the performance of the D&S algorithm on the MDM data.

The metric used to evaluate the performance is the same as used in the NIST RT04s and RT04f evaluations. An optimum one-to-one mapping of reference speaker ID to system output ID is performed and the regions of erroneous or missed mapping are considered errors. A collar of 0.25s is allowed at the speaker changes. As in the RT04s evaluation, two variants are considered: errors including overlapping speech and errors excluding overlapping regions.

In Tables 1 and 2 we show the results for the MDM and SDM tasks in the RT04s Meetings database (evaluation and development sets). Note that the "ALL" results are not the arithmetic mean of the individual scores; they are a time-weighted combination of these scores.

Evaluation set	#	non overlap		overlap	
	mics	MDM	SDM	MDM	SDM
CMU_20030109-1530	1	10.91	10.91	31.08	31.08
CMU_20030109-1600	1	11.10	11.10	31.10	31.10
ICSI_20000807-1000	6	8.32	22.17	21.12	34.01
ICSI_20011030-1030	6	14.56	27.23	39.54	47.28
LDC_20011121-1700	10	6.16	16.59	21.97	30.73
LDC_20011207-1800	4	37.14	43.81	46.36	51.26
NIST_20030623-1409	7	2.19	7.53	17.50	22.65
NIST_20030925-1517	7	42.69	44.35	51.87	52.33
ALL		**15.75**	22.22	**32.12**	37.07

Table 1: *DER for the RT04s Meetings Database, eval set*

Development set	#	non overlap		overlap	
	mics	MDM	SDM	MDM	SDM
CMU_20020319-1400	1	50.36	50.36	52.39	52.39
CMU_20020320-1500	1	47.97	47.97	49.14	49.14
ICSI_20010208-1430	6	6.47	22.74	10.88	26.18
ICSI_20010322-1450	6	6.47	13.59	15.13	21.46
LDC_20011116-1400	7	9.86	4.13	14.98	10.10
LDC_20011116-1500	8	9.57	17.28	25.38	30.38
NIST_20020214-1148	7	34.04	37.20	37.80	40.96
NIST_20020305-1007	6	17.72	32.50	22.78	36.27
ALL		**22.73**	27.95	**28.42**	33.04

Table 2: *DER for the RT04s Meetings Database, devel set*

In both data sets we see an improvement when using D&S to enhance the signal. Improvements between 13.7% and 29.11% relative are obtained (depending on the set) between the MDM and SDM cases. In both sets the CMU meetings only had 1 microphone, therefore the MDM and SDM results are

identical. Not taking into account the CMU shows, the relative improvements range between 17.1% and 33%.

No special treatment was given to microphone types in any of these experiments. Therefore, even microphones with much worse SNR (e.g. the mock-PDA microphones in the ICSI corpus) than the centrally located mics were used in the D&S.

In order to study the importance of an accurate TDOA estimation, an enhanced signal has been constructed by artificially setting the TDOA = 0. The signal constructed using TDOA = 0 was then fed into the ICSI-SRI diarization system, obtaining the results shown in table 3.

System	RT04s eval		RT04s devel	
	Non Ovl.	Ovl.	Non Ovl.	Ovl.
SDM system	22.22	37.07	27.95	33.04
0 delay system	18.72	34.27	29.16	34.16
D&S system	15.75	32.12	22.73	28.42

Table 3: *Comparison between SDM, MDM and MDM with 0 delay*

In general all meetings improve their diarization performance compared to the SDM when we combine all channels. In the development set the show LDC_20011116-1500 contains several sound artifacts in some channels that seem to affect much more the resulting signal when they are not properly time-aligned. That lead to a DER of 53.04% - 55.69%, which is why we obtain an overall worse score for the TDOA=0 version versus the SDM.

For comparison purposes, in Table 4 we show the official results from NIST for the best submissions to the RT04s evaluation, compared with the results of our proposed system.

RT04f eval MDM	Non Overlap	Overlap
LIA-CLIPS	22.79	37.22
ISL	28.17	40.19
Macquarie Univ.	62.02	69.09
Proposed system	**15.75**	**32.12**

Table 4: *Comparison between RT04s Meetings Diarizarion MDM task and our proposed system using D&S. Note that the results for the official submissions were taken directly from the official NIST RT04s web site and it is likely that these sites have significantly improved their systems since the time of the official evaluation.*

Our proposed system using D&S shows improvements of 30.8%(MDM) and 13.70%(SDM) relative compared to the best performing systems in official RT04s evaluation. We believe this is due both to the use of D&S to enhance the signal, and also to the robustness of the speaker diarization system we are using. Even our SDM DER (i.e. with no D&S) shows improvement over the best official MDM results.

The system presented is very fast as it only involves a delay and sum step (running at an average of 0.3 times real-time) and the diarization of only one channel. While the best system in RT04s eval runs a full diarization on each channel twice, and then post-processes the results.

The results obtained using this D&S technique on meetings are comparable with the results we reported using the same diarization system in the RT04f Broadcast News Diarizarion evaluation (17.91% DER) (see [5]). We believe this is an indication of the robustness of the diarization system given a change of task.

5. Conclusion

In this paper we present a system for speaker diarization in the meetings environment. Our system exploits the existence of multiple channels to obtain an enhanced signal using the delay-and-sum algorithm. The input signal from the different channels is analyzed with a small sliding window and the TDOA values are estimated and treated for continuity. The time-delayed signals are summed to obtain the resulting enhanced signal. We use the ICSI-SRI Broadcast News diarization system on this signal to obtain the diarization result. Tests on the official RT04s databases show an improvement compared to the use of only a single channel and also show improvements over the best official results from the evaluation.

6. Acknowledgements

This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-Party Interaction, FP6-506811, publication AMI-15). I would also like to thank Marc Ferras, Barbara Peskin, Nikki Mirghafori and James Fung for so many helpful discussions.

7. References

[1] Nist rich transcription evaluations. [Online]. Available: http://www.nist.gov/speech/tests/rt

[2] S. Cassidy, "The macquarie speaker diarization system for rt04s," in *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada, 2004.

[3] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, "Speaker segmentation and clustering in meetings," in *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada, 2004.

[4] C. Fredouille, D. Moraru, S. Meignier, L. Besacier, and J.-F. Bonastre, "The nist 2004 spring rich transcription evaluation: Two-axis merging strategy in the context of multiple distant microphone based meeting speaker segmentation," in *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada, 2004.

[5] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system," in *Rich Transcription Workshop*, New Jersey, USA, 2004.

[6] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," Atlanta, USA, May 1996.

[7] J. Flanagan, J. Johnson, R. Kahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of the Acoustic Society of America*, vol. 78, pp. 1508–1518, November 1994.

[8] N. Mirghafori, A. Stolcke, C. Wooters, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf, "From switchboard to meetings: Development of the 2004 icsi-sri-uw meeting recognition system," in *ICSLP-04*, Jeju Island, Korea, October 2004.