SPEAKER DIARIZATION FOR MULTI-PARTY 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

One of the sub-tasks of the Spring 2004 and Spring 2005 NIST Meetings evaluations requires segmenting multi-party meetings into speaker-homogeneous regions using data from multiple distant microphones (the "MDM" sub-task). One approach to this task is to run a speaker segmentation system on each of the microphone channels separately, and then merge the results. This can be thought of as a many-to-one *post-processing* approach. In this paper we propose an alternative approach in which we use delay-and-sum beamforming techniques to fuse the signals from each of the multiple distant microphones into a single enhanced signal. This approach can be thought of a many-to-one *pre-processing* approach. In the pre-processing approach we propose, the time delay of arrival (TDOA) between each of the multiple distant channels and a reference channel is computed incrementally using a window that steps through the signals from each of the multiple microphones. No information about the locations or setup of the microphones is required. Using the TDOA information, the channels are first aligned and then summed and the resulting "enhanced" signal is clustered using our standard speaker diarization system. We test our approach on the 2004 and 2005 NIST meetings evaluation databases and show that the technique performs very well.

1. INTRODUCTION

Speaker diarization attempts to answer the question "Who spoke when?" in a multi-person recording. In recent years there has been extensive research in speaker diarization for the Broadcast News (BN) environment ([1]). In this environment, single channel recordings from radio and TV programs include speech (in various acoustic environments), music, advertisements and other background noises. More recently, research has begun in the area of multi-party meetings where the speaker diarization task has 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, few publications have addressed the problem. The baseline approach is to consider only the best channel (usually from the most centrally located microphone) and perform speaker diarization on it. This was 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 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 process all the channels and then post-process the resulting segmentations. In [4] an iterative process is used looking for the longest speaker intervention from all channels.

In this paper we present a signal processing approach where a classic channel weighted delay-and-sum (D&S) is used to combine all channels into one single enhanced channel that is then clustered using the ICSI-SRI speaker diarization system ([5], [6]). The reference channel is selected automatically using a Signal-to-Noise Ratio (SNR)

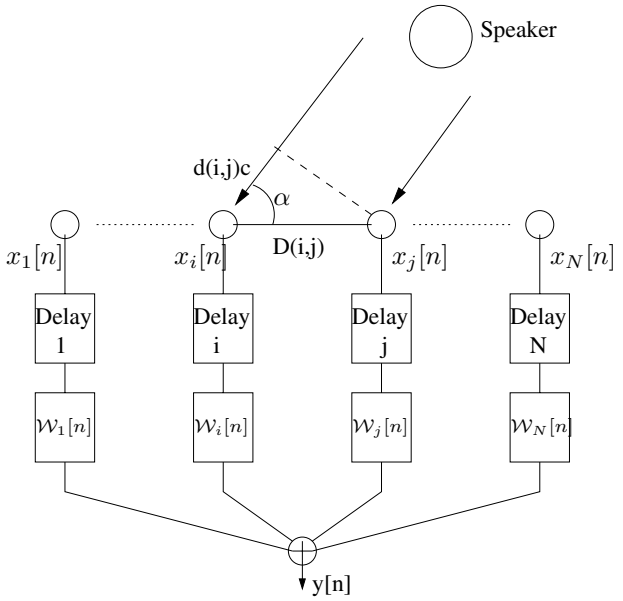

Fig. 1. *Delay-and-sum system*

comparison. The weight for each channel is adapted based on the correlation between that channel and the reference channel. By using the D&S algorithm along a moving window, we attempt to align all the channels with respect to the speaker who is currently talking. This has the effect of improving the overall quality of the resulting signal 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 ([7]). Two 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 reviewed. 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 ([8]) is a simple yet effective method 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:

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

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 non-dispersive wave propagation, we obtain the delay (in # of samples) as

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

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.).

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

$$y[n] = \mathcal{W}_0[n] \cdot x_0[n] + \sum_{i=1}^{N-1} \mathcal{W}_i[n] \cdot x_i[n - d(0,i)] \quad (3)$$

where each channel is weighted with $\mathcal{W}_i[n]$, which can be constant or variable in time. The basic delay&sum systems use $\mathcal{W}_i = \frac{1}{N}$.

By adding together the time-aligned signals, the speech segments get enhanced and the noise (assuming it is random with similar properties) is minimized. Using the D&S we can obtain (according to [8]) up to a 3db SNR improvement each time that the number of microphones double.

2.1. TDOA Estimation via GCC-PHAT

In order to estimate the TDOA between two segments from two microphones we cannot use Eq. 2 because in speaker diarization we do not know the number of speakers or their locations. Therefore we use a modified version of the Generalized Cross Correlation (GCC) called "generalized cross correlation with phase transform" (GCC-PHAT) (see [7]).

Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is defined as:

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

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:

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

Where $\hat{R}_{PHAT}(d)$ is the inverse Fourier transform of Eq. 4. 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.

2.2. Reference Channel Selection

In order to run the delay&sum algorithm on a set of microphones we need to define a particular channel as the reference (channel 0 in Eq. 3). Within the Rich Transcription Meetings Evaluation, one channel is defined as the most central in the meeting room (this is the one selected to run the Single Distant Microphone (SDM) task).

We have found that although the SDM channel might be the most central mic, the nature and structure of meetings (and different microphones specs) causes a different channel to sometimes perform better as reference channel for the delay&sum processing. This is provably because the optimum acoustic center doesn't match the physical center.

We use the Signal-to-Noise Ratio (SNR) to select the best channel to act as reference channel. To estimate the SNR value we use a tool provided by Prof. Hans-Gunter Hirsch (Niederrhein University, Germany) which performs a 2-step process:

1. Stationary segments are detected based on a Mel frequency analysis using the short term subband energies for all subbands. As soon as the subband energy exceeds a certain threshold (defined as the average of the previous energies) this is considered a possible indication for the presence of speech. When a certain number of subbands exceed the threshold it indicates the start of a speech segment. Similar thresholding is used to determine the transition from speech to non-speech.

2. The SNR is computed as $10log_{10}(\frac{S}{N})$ where N is the RMS value of the non-speech parts and S is obtained from the RMS of the speech parts, considering that they are $X = S + N$. Such energy is computed over the "A" filtered data.

More information regarding how the SNR is obtained can be found in [9].

2.3. Individual Channel Weighting Using Correlations

In the formulation of the delay&sum processing, the additive noise components on each of the channels are expected to be random processes with very similar probability distributions. This allows the noise on each channel to be minimized when the delay-adjusted channels are summed. In standard beamforming systems, this noise cancellation is achieved through the use of identical microphones placed only a few inches apart.

In the meetings room we assume that all of the distant microphones form a microphone array. However, having different types of microphones changes the characteristics of the signal being recorded and therefore changes the probability distributions of the resulting additive noise. Also when two microphones are far from each other, the speech they record will be affected by noise of a different nature, due to the room's impulse response, and will have different quality depending on the position of the speaker talking.

We address this issue by weighting each channel in the delay&sum processing. The weights are adapted continuously during the meeting. This is inspired by the fact that the different channels will have different qualities depending on their relative distance to the person speaking, which can change constantly during a recording.

The weight for channel i at step n ($\mathcal{W}_i[n]$) is computed in the following way:

$$\mathcal{W}_i[n] = \begin{cases} \frac{1}{\#\text{Channels}} & n = 0 \\ (1-\alpha)\cdot\mathcal{W}_i[n-1] + & \\ \alpha\cdot\text{xcorr}(i,\text{ref}) & \text{otherwise} \end{cases} \quad (6)$$

where $xcorr(i, ref)$ is the cross-correlation between the delay-adjusted segment for channel i and the reference channel. When i=reference, it is just the power of the reference channel. If the cross-correlation becomes negative, it is set to 0.0. We empirically set $\alpha = 0.05$.

3. SYSTEM IMPLEMENTATION

Figure 2 presents the basic blocks forming the system presented in this paper. The raw signal coming from the available channels is individually Wiener-filtered to improve the SNR in the same way as was done in the ICSI-SRI-UW Meetings recognition system ([10]). Then the channel weighted D&S is performed using a scrolling window of 500ms with 50% overlap.

The selection of a 500ms window constitutes a tradeoff. The lower bound is set by the minimum number of samples needed to accurately estimate the correlation. The upper bound is defined by the accuracy desired in obtaining the correct TDOA. Experimentally, we have found a value of 500ms to be a good tradeoff for both issues and for system speed.

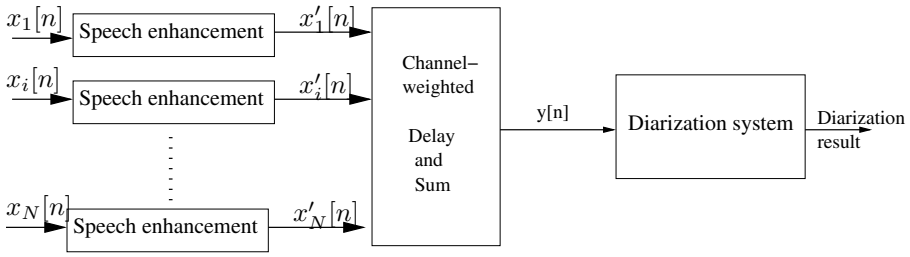

Fig. 2. *Diarization system for meetings data*

3.1. Robust TDOA Estimation

To obtain the Time Delay of Arrival (TDOA) for each segment of a channel, the GCC-PHAT ([7]) is computed between the segment and the corresponding segment in the reference channel. 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. To obtain the TDOA value we first search for the 8 major peak values within a region of $\pm$20ms (360 samples) allowable delay. This defines a maximum distance between both channels of 7 meters. Since our ultimate goal is to enhance the output signal (not to obtain an optimum estimate of the position of each speaker through their TDOA), the following smoothing filters are applied to the computed maxima in order to find the desired TDOA:

1. TDOA Continuity: In the presence of multiple speakers or impulsive noises, the main peak of the GCC-PHAT may shift to a new source location, or jump back and forth between sources. This filtering step tries to enforce continuity of the TDOA values by searching for a $d(n) = d(n-1) \pm \Delta$ within all computed maxima. In our system we use $\Delta = 10$ samples.

2. Confidence level: The TDOA estimation in silence regions or in low SNR regions is not reliable and often is completely wrong. Thus, we threshold the max GCC-PHAT value using a cutoff of 0.1. When the max GCC-PHAT value doesn't exceed this threshold, we force continuity by setting $d(n) = d(n-1)$.

The resulting TDOA estimates are applied to each windowed segment. The cross-correlations between the segments from each channel and the reference are then computed and the weights adapted. The segments are then summed together using these 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 is then analyzed by our speech/non-speech detector in order to eliminate all non-speech regions from the diarization process. The resulting segments are then processed using the ICSI-SRI speaker diarization system. This system is a bottom-up agglomerative system that uses a BIC-like measure as a merging and stopping criteria. The algorithm is initialized with 10 clusters of equal size from the speech-only enhanced input signal and iteratively performs a re-segmentation and cluster merging process until the stopping criterion is met. This system does not use any pre-trained models or threshold adjustments. Thus porting it to a new task or new data is straightforward. For a more detailed account of the diarization system, see [5]; [11].

4. EXPERIMENTS AND RESULTS

Speaker diarization experiments were conducted using the data distributed for the NIST Rich Transcription 2004 and 2005 Spring Meeting Recognition Evaluation, RT04s and RT05s ([1]). This consists of excerpts from multi-party meetings in English collected at six different sites. From each meeting only an excerpt of 10 to 12 minutes is evaluated. The number of distant microphones available varies among the meetings, ranging from one microphone (in RT04 CMU recordings) to 16 (in one AMI recording). We show results for two conditions: MDM (Multiple Distant Microphones), where all microphones can be used, and SDM (Single Distant Microphones), where only the most centrally located microphone (as defined by NIST) is used. We use the SDM condition as a baseline to compare the performance of the D&S algorithm on the MDM data.

The metric used to evaluate the performance of the system is the same as is used in the NIST RT evaluations. An optimal one-to-one mapping of reference speaker ID to system output ID is performed and the error is computed as the percentage of time that the system assigns the wrong speaker label.

Tables 2, 1 and 3 show the Diarization Error Rates (DERs) for the MDM and SDM tasks on the RT04s and

RT05s Meetings database. Note that the "ALL" results are not the arithmetic mean of the individual scores; they are a time-weighted combination of these scores.

Development set RT04s	# mics	non overlap SDM	non overlap MDM
CMU_20020319-1400	1	32.17	32.17
CMU_20020320-1500	1	16.98	16.98
ICSI_20010208-1430	6	13.18	10.65
ICSI_20010322-1450	6	19.26	14.26
LDC_20011116-1400	7	6.31	3.55
LDC_20011116-1500	8	7.14	11.96
NIST_20020214-1148	7	38.32	24.04
NIST_20020305-1007	6	33.24	15.93
ALL		21.32	**16.22**

Table 1. *DER for the RT04s Meetings Database, devel set*

Evaluation set RT04s	# mics	non overlap SDM	non overlap MDM
CMU_20030109-1530	1	26.78	26.78
CMU_20030109-1600	1	15.66	15.66
ICSI_20000807-1000	6	11.12	7.73
ICSI_20011030-1030	6	13.62	26.81
LDC_20011121-1700	10	2.54	3.36
LDC_20011207-1800	4	33.46	27.85
NIST_20030623-1409	7	5.04	4.18
NIST_20030925-1517	7	35.95	34.75
ALL		17.32	**16.92**

Table 2. *DER for the RT04s Meetings Database, eval set*

Evaluation set RT04s	# mics	non overlap SDM	non overlap MDM
AMI_20041210-1052	12	10.51	10.64
AMI_20050204-1206	16	10.44	9.54
CMU_20050228-1615	3	15.00	7.77
CMU_20050301-1415	3	10.59	17.30
ICSI_20010531-1030	6	14.30	13.49
ICSI_20011113-1100	6	24.43	30.86
NIST_20050412-1303	7	10.84	7.88
NIST_20050427-0939	7	12.70	10.33
VT_20050304-1300	2	9.31	7.04
VT_20050318-1430	2	42.19	39.45
ALL		15.33	**14.81**

Table 3. *DER for the RT05s Meetings Database, eval set*

We observe that in the three evaluated sets, the MDM system outperforms the SDM channel, with the RT04s devel set showing the most improvement (24% relative). For each set, we show each meeting's individual scores to see the contribution of each meeting to the overall performance of the systems. All CMU meetings on RT04s evaluations only provided one channel; therefore no signal enhancement was possible and the scores are identical for both cases.

There seems not to be any direct relationship between the number of available channels and the improvement seen between SDM and MDM tasks. This is probably due to the differences between the different microphones within a meeting (as discussed in section 2.3). This might be the reason why in two ICSI meetings SDM outperforms MDM. These meetings contain two low-quality microphones attached to a mock PDA. Some of the channels in the other meetings with the same performance have artifacts that degrade the resulting signal and therefore can affect the final diarization result.

In order to break down the improvement due to some of the individual modules, we show in table 4 the overall DER for the SDM, the full system MDM cases and three intermediate systems.

System	weights	RT04s eval	RT04s devel	RT05s eval
SDM	n/a	17.32	21.32	15.33
D&S MDM	equal	18.75	15.62	16.33
0-D&S MDM	SNR+corr	19.63	18.30	17.23
D&S MDM	corr	20.14	17.46	16.16
D&S MDM	SNR+corr	16.92	16.22	14.81

Table 4. *Comparison between different configurations (non-overlap)*

The first and fifth rows contain the results presented above, for the SDM and MDM systems (shown in tables 1, 2 and 3). The second row (system D&S MDM with equal weights) shows a basic delay&sum system, where the weight for each channel is set to $\frac{1}{N}$. Both eval sets obtain worse results than the baseline SDM system. The third row (system 0-D&S MDM with SNR+corr weights) presents a system that doesn't perform any TDOA estimation, adding all channels with delay 0. We see how the RT04s devel set still obtains an improvement, but the other two sets don't. Finally, the fourth (D&S MDM with corr weights) system row uses a weighted delay&sum system using the SDM channel as the reference channel (versus using the channel with the best SNR as the reference). As in the previous cases, only the RT04s devel set outperform the baseline.

Although we didn't participate on the RT04s Meetings Diarization Evaluation, we compared the performance of the Evaluation results for the official submissions as reported on the NIST RT04s web site, [1]. Our system outperforms the best presented system by 25% relative. On the

just-completed RT05s Meetings evaluation, our system also performed well, as can be seen by the DER obtained in the RT05s eval set.

The system presented here is computationally efficient 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, 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 indicates that the core system (used for Diarization in meetings and Broadcast News) is robust to a change of task and data.

5. CONCLUSION

In this paper we presented 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 then adjusted for continuity. The time-delayed signals are individually weighted and summed to obtain an enhanced single-channel signal. We then used the ICSI-SRI diarization system on this signal to perform speaker diarization. Tests on the official RT04s and RT05s databases show an improvement compared to the use of only a single channel.

By combining multiple channels into a single enhanced channel, we are essentially ignoring valuable information e.g. which channel has the highest energy at a given point in time. Thus, in future work, we will try to improve the delay-and-sum processing by using extra information extracted during processing (e.g. TDOA values, correlation weights, relative energy between microphones, etc.).

6. ACKNOWLEDGEMENTS

We would like to acknowledge Hans-Guenter Hirsch for his help with the SNR estimation system. This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-Party Interaction, FP6-506811). We would also like to thank Marc Ferras, Barbara Peskin, Nikki Mirghafori and James Fung for 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] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *ASRU'03*, US Virgin Islands, USA, Dec. 2003.

[7] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *ICASSP-97*, Munich, Germany, 1997.

[8] 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.

[9] H.-G. Hirsch, "HMM adaptation for applications in telecommunication," *Speech Communication*, no. 34, pp. 127–139, 2001.

[10] 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.

[11] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Rich Transcription 2005 Spring Meeting Recognition Evaluation (to appear)*, Edinburgh, Great Brittain, July 2005.