Acoustic Beamforming for Speaker Diarization of Meetings

Xavier Anguera, *Member, IEEE,* Chuck Wooters, *Member, IEEE,* Javier Hernando, *Member, IEEE*

***Abstract*—When performing speaker diarization on recordings from meetings, multiple microphones of different qualities are usually available and distributed around the meeting room. Although several approaches have been proposed in recent years to take advantage of multiple microphones, they are either too computationally expensive and not easily scalable or they can not outperform the simpler case of using the best single microphone. In this work the use of classic acoustic beamforming techniques is proposed together with several novel algorithms to create a complete frontend for speaker diarization in the meeting room domain. New techniques we are present include blind reference-channel selection, two-step Time Delay of Arrival (TDOA) Viterbi postprocessing, and a dynamic output signal weighting algorithm, together with using such TDOA values in the diarization to complement the acoustic information. Tests on speaker diarization show a 25% relative improvement on the test set compared to using a single most centrally located microphone. Additional experimental results show improvements using these techniques in a speech recognition task.**

***Index Terms*—acoustic beamforming, speaker diarization, speaker segmentation and clustering, meetings processing.**

I. Introduction

Possibly the most noticeable difference when performing speaker diarization in the meetings environment versus other domains (like broadcast news or telephone speech) is the availability, at times, of multiple microphone channels, synchronously recording what occurs in the meeting. Their varied locations, quantity, and wide range of signal quality has made it difficult to come up with automatic ways to take advantage of these multiple channels for speech-related tasks such as speaker diarization.

In the system developed by Macquarie University [1] and the TNO/AMI systems ([2] and [3]), either the most centrally located microphone (known a priori) or a randomly selected single microphone was used for speaker diarization. This approach was designed to prevent low quality microphones from affecting the results. Such approaches ignore the potential advantage of using multiple microphones- making use of the alternate microphone channels to create an improved signal as the interaction moves from one speaker to another. Several alternatives have been proposed to analyze and switch channels dynamically as the meeting progresses. At CMU [4] this is done before any speaker diarization processing by using a combination of energy and signal-to-noise metrics. However, this approach creates a patchwork-type signal which could make interfere with the speaker diarization algorithms. In an alternative presented in an initial LIA implementation [5], all channels were processed in parallel and the best segments from each channel were selected at the output. This technique is computationally expensive as a full speaker diarization processing must be performed for every channel. Later, LIA proposed ([6] and [7]) a weighted sum of all channels into a single channel prior to performing diarization. However, this approach does not take into account the fact that the signals may be misaligned due to the propagation time of speech through the air or hardware timing issues, resulting in a summed signal that contains echoes, and usually performs worse than the best single channel.

To take advantage of the multiple microphones available in a typical meeting room, we previously proposed ([8] and [9]) the use of microphone array beamforming for speech/acoustic enhancement (see [10], [11]). Although the task at hand differs from the classic due to some of the assumptions in the beamforming theory, it was found to be beneficial to use it as a starting-point for taking advantage of the multiple distant microphones.

In this work we propose a full acoustic beamforming frontend, based on weighted-delay&sum techniques [10], aimed at creating a single enhanced signal from an unknown number of multiple microphone channels. This system is designed for recordings made in meetings in which several speakers and other sources of interference are present. Several new algorithms are proposed to adapt the general beamforming theory to this particular domain. Algorithms proposed include the automatic selection of the reference channel, the computation of the N-best channel delays, postprocessing techniques to select the optimum delay values (including a noise thresholding and a two-step selection algorithm via Viterbi decoding), and a dynamic channel-weight estimation to reduce the negative impact of low quality channels.

The system presented here was used as part of ICSI's submission to the Spring 2006 Rich Transcription evaluation (RT06s) organized by NIST [12], both in the speaker diarization and in the speech recognition systems. Additionally, the software is currently available as open-source [13].

The next section describes the modules used in the acoustic beamforming system. Then, we present experimental results showing the improvements gained by using the new system within the task of speaker diarization, and finally, we present results for the task of speech recognition.

At the time of this work X. Anguera was visiting the International Computer Science Institute (ICSI), Berkeley, California. C. Wooters is currently with ICSI and J. Hernando is with Universitat Politecnica de Catalunya (UPC), Barcelona, Spain.

II. Multichannel Acoustic Beamforming System Implementation

The acoustic beamforming system is based on the weighted-delay&sum microphone array theory, which is a generalization of the well known weighted-delay&sum beamforming technique ([14], [15]). The signal output $y[n]$ is expressed as the weighted sum of the different channels as follows:

$$y[n] = \sum_{m=1}^{M} W_m[n] x_m[n - \mathrm{TDOA}^{(m,\mathrm{ref})}[n]] \qquad (1)$$

where $W_m[n]$ is the relative weight for microphone m (out of M microphones) at instant n, with the sum of all weights equals to 1; $x_m[n]$ is the signal for each channel, and $\mathrm{TDOA}^{(m,\mathrm{ref})}[n]$ (Time Delay of Arrival) is the relative delay between each channel and the reference channel, in order to obtain all signals aligned with each other at each instant n. In practice, $\mathrm{TDOA}^{(m,\mathrm{ref})}[n]$ is estimated via cross correlation techniques once every several acoustic frames. In the implementation presented here it corresponds to once every 250 ms, using GCC-PHAT (Generalized Cross Correlation with Phase Transform) as proposed in [16] and [17] and described below. We will refer to these as "acoustic segments", and we will refer to the (usually larger) set of frames used to estimate the cross correlation measure the "analysis window".

The weighted-delay&sum technique was selected for use in the meetings domain given the following set of constraints:

- Unknown locations of the microphones in the meeting room.
- Non-uniform microphone settings (gain, recording offsets, etc.)
- Unknown location and number of speakers in the room. Due to this constraint, any techniques based on known source locations are unsuitable.
- Unknown number of microphones in the meeting room. The system should be able to handle from 2 to >100 microphone channels.

Figure 1 shows the different blocks involved in the proposed weighted-delay&sum process. The process can be split into four main blocks. First, signal enhancement via Wiener filtering is performed on each individual channel to reduce the noise. Next, the information extraction block is in charge of estimating which channel to use as the reference channel, an overall weighting factor for the output signal, the skew present in the ICSI meetings, and the N-best TDOA values at each analysis segment. Third, a selection of the appropriate TDOA delays between signals is obtained in order to optimally align the channels before the sum. Finally, the signals are aligned and summed. The output of the system is composed of the acoustic signal and a vector of TDOA values, which can be used as extra information about a speaker's position. A more detailed description of each block follows.

A. Individual Channel Signal Enhancement

Prior to doing any multichannel beamforming, each individual channel is Wiener filtered [18]. This aims at cleaning the signal of corrupting noise, which is assumed to be additive and of a stochastic nature. The implementation of Wiener filtering is taken from the ICSI-SRI-UW system used for ASR in [19], and applied to each channel independently. This implementation performs an internal speech/non-speech and noise power estimation for each channel independently, ignoring any multichannel properties or microphone locations. The use of such filtering improves the beamforming as it increases the quality of the signal, even though it introduces a small phase nonlinearity given that the filter is not of linear phase. Alternative multichannel Wiener filters were not considered but could further improve results by taking advantage of redundancies in the different input channels.

B. Meeting Information Extraction block

The algorithms in this block extract information from the input signals to be used further on in the process to construct the output signal. It is composed of four algorithms- reference channel estimation, overall channels weighting factor, ICSI meetings skew estimation, and the TDOA N-best delays estimation.

1) Reference Channel Estimation: This algorithm attempts to automatically find the most centrally located and best quality channel to be used as the reference channel in further processing. It is important for this channel to be the best representative of the acoustics in the meeting, as the correct estimation of the delays of each of the channels depends on the reference chosen.

In the meetings used for the Rich Transcription evaluations [20], there is one microphone that is selected as the most centrally located microphone. This microphone channel is used in the Single Distant Microphone (SDM) task. The SDM channel is chosen given the room layout and the prior knowledge of the microphone types. This module presented here, ignores that channel chosen for the SDM condition and selects one microphone automatically based only on the acoustics. This is intended for system robustness in cases where absolutely no information is available on the room layout or microphone placements.

In order to find the reference channel, we use a metric based on a time-average of the cross-correlation between each channel i and all of the others $j = 1 \dots M, j \neq i$, computed on segments of 1 second, as

$$\overline{\mathrm{xcorr}}_i = \frac{1}{K(M-1)} \sum_{k=1}^{K} \sum_{j=1, j\neq i}^{M} \mathrm{xcorr}[i, j;\ k] \qquad (2)$$

where M is the total number of channels/microphones and $K = 200$ indicates the number of one second blocks used in the average. The $\mathrm{xcorr}[i, j;\ k]$ indicates a standard cross-correlation measure between channels i and j for each block k. The channel i with the highest average cross-correlation was chosen as the reference channel. An alternative SNR metric was analyzed and the results were not conclusive as to which method performed better in all cases. The cross-correlation metric was chosen as it matches the algorithm search for maximum correlation values and because it is simple to implement.

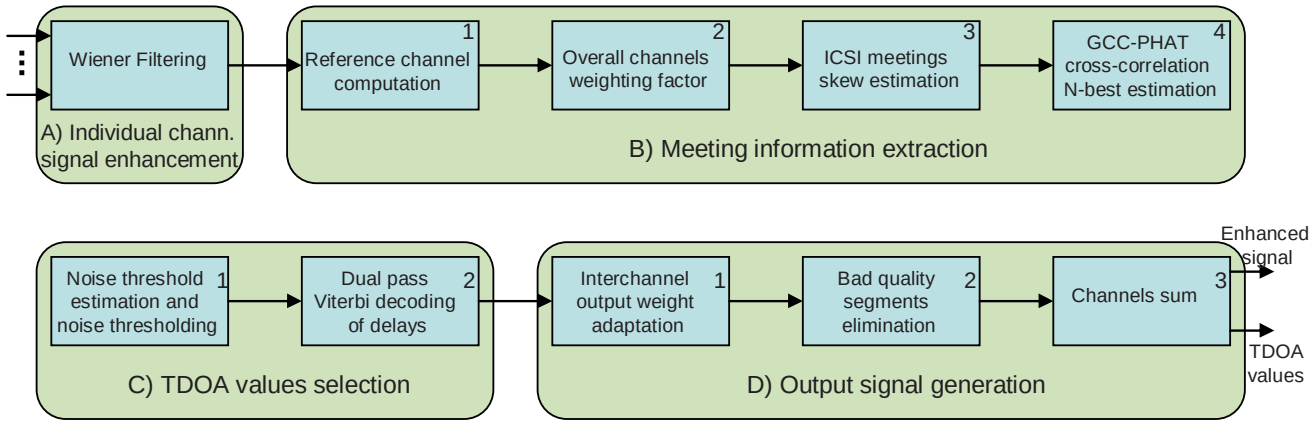


Fig. 1. *Weighted-delay&sum block diagram*

2) Overall Channels Weighting Factor: For practical reasons, speech processing applications use acoustic data that was sampled with a limited number of bits (e.g. 16 bits per sample) providing a certain amount of dynamic range, which is often not fully used because the recorded signals are of low amplitude. When summing up several input signals, we are increasing the resolution in the resulting signal, and thus we must try to take advantage of as much of the output resolution as possible. The overall channel weighting factor is used to normalize the input signals to match the file's available dynamic range. It is useful for low amplitude input signals since the beamformed output has greater resolution and therefore can be scaled appropriately to minimize the quantization errors generated by scaling it to the output sampling requirements.

There are several methods in signal processing for finding the maximum value of a signal in order to perform amplitude normalization. These include- compute the absolute maximum amplitude, the Root Mean Square (RMS) value, or other variations of it, over the entire recording. It was observed in meetings data that the signals may contain low energy areas (silence regions) with short average durations, and high energy areas (impulsive noises like door slams, or laughs), with even shorter duration. Using the absolute maximum or RMS would "saturate" the normalizing factor to the highest possible value or bias it according to the amount of silence in the meeting. So instead, we chose a windowed maximum averaging to try to increase the likelihood that every window contains some speech. In each window the maximum value is found and these max values are averaged over the entire recording. The weighting factor was obtained directly from this average.

3) ICSI Meetings Skew Estimation: This module was created to deal with the meetings that come from the ICSI Meeting Corpus, some of which have an error in the synchronization of the channels. This was originally detected and reported in [21], indicating that the hardware used for the recordings was found not to keep an exact synchronization between the different channels, resulting in a skew between channels of multiples of 2.64 ms. It is not possible to know beforehand the amount of skew of each of the channels as the room setup did not follow a consistent ordering regarding the connections to the hardware being used. Therefore we need to automatically detect such skew so that it does not affect the beamforming.

The artificially generated skew does not affect the general processing of the channels by an ASR system as it does not need exact time alignment between the channels- utterance boundaries always include a silence "guard" region, and the usual parametrizations (10-20ms long) cover small time differences.

It does pose a problem though when computing the delays between channels as it introduces an artificial delay between channel pairs, which forces us to use a larger analysis window for the ICSI meetings than with other meetings in order to compute the delays accurately. This increases the chance of delay estimation error. This module is therefore used to estimate the skew between each channel and the reference channel (in the case of ICSI meetings) and use it as a constant bias in the rest of the delay processing.

In order to estimate the bias, an average cross-correlation metric was put in place in order to obtain the average (across time) delay between each channel and the reference channel for a set of long acoustic windows (around 20 seconds), evenly distributed along the meeting.

4) TDOA N-best delays estimation: The computation of the time delay of arrival (TDOA) between each of the channels and the reference channel is computed in segments of 250 ms. This allows the beamforming to quickly modify its beam steering whenever the active speaker changes. In this implementation the TDOA was computed over a window of 500ms (called the analysis window), which covers the current analysis segment and the next. The size of the analysis window and of the segment size constitute a tradeoff. A large analysis window or segment window leads to a reduction in the resolution of changes in the TDOA. On the other hand, using a small analysis window reduces the robustness of the estimation. The reduction of the segment size also increases the computational cost of the system, while not increasing the quality of the output signal. The selection of the scroll and analysis window sizes was done empirically given some development data and no exhaustive study was performed to fine-tune these values.

In order to compute the TDOA between the reference channel and any other channel for any given segment it is usual to estimate it as the delay that maximizes the cross-correlation between the two segments. In current beamforming systems, the use of the cross-correlation in its classical form ($R^{i,\mathrm{ref}}_{\mathrm{xcorr}}(d) = \sum_{n=0}^{N} x_i[n]x_{\mathrm{ref}}[n+d]$) is avoided as it is very sensitive to noise and reverberation. To improve robustness against these problems, it is common practice to use the GCC-PHAT. Such variation to the standard cross-correlation proposes an amplitude normalization in the frequency domain, maintaining the phase information, which conveys the delay information between the signals.

Given two signals $x_i(n)$ and $x_{\mathrm{ref}}(n)$, the GCC-PHAT is computed with:

$$\hat{R}^{i,\mathrm{ref}}_{\mathrm{PHAT}}(d) = \mathcal{F}^{-1}\Big(\frac{X_i(f)[X_{\mathrm{ref}}(f)]^*}{|X_i(f)[X_{\mathrm{ref}}(f)]^*|}\Big) \tag{3}$$

Where $X_i(f)$ and $X_{\mathrm{ref}}(f)$ are the Fourier transforms of the two signals, $\mathcal{F}^{-1}$ indicates the inverse Fourier transformation, $[\ ]^*$ denotes the complex conjugate and $|\cdot|$ is the modulus. The resulting $\hat{R}^{i,\mathrm{ref}}_{\mathrm{PHAT}}(d)$ is the correlation function between signals i and ref. All possible values range from 0 to 1 given the frequency domain amplitude normalization performed.

The Time delay of arrival (TDOA) for these two microphones (i and ref) is estimated as

$$\mathrm{TDOA}^i_1 = \arg\max_d \big(\hat{R}^{i,\mathrm{ref}}_{\mathrm{PHAT}}(d)\big) \tag{4}$$

which we noted with subscript 1 (1st-best) to differentiate it from further computed values.

Although the maximum value of $\hat{R}^{i,\mathrm{ref}}_{\mathrm{PHAT}}(d)$ corresponds to the estimated TDOA for that particular segment and microphones pair, it does not always "point" at the correct speaker during that segment. In the system proposed here the top N relative maxima of $\hat{R}^{i,\mathrm{ref}}_{\mathrm{PHAT}}(d)$ are computed instead (we use N around 4), and several post-processing techniques are used to "stabilize" and choose the appropriate delay before aligning the signals for the sum. Therefore, for each analysis segment we obtain a vector TDOA^i_n for microphone i with $m = 1 \ldots M, i \neq \mathrm{ref}$ with its corresponding correlation values GCC-PHAT^i_n with $n = 1 \ldots N$.

We could isolate three cases where it was considered not appropriate to use the absolute maximum (1st-best) from $\hat{R}^{i,\mathrm{ref}}_{\mathrm{PHAT}}(d)$. On the one hand, the maximum can be due to spurious noises or events not related to the active speaker, and the active speaker is actually represented by another local maximum of the cross-correlation. On the other hand, when two or more speakers are speaking simultaneously, each speaker will be represented by a different maximum in the cross-correlation function, but the absolute maximum might not be constantly assigned to the same speaker resulting in artificial speaker switching. Finally, when the segment that has been processed is entirely filled with non-speech acoustic data (either noise or random acoustic events) the $\hat{R}^{i,\mathrm{ref}}_{\mathrm{PHAT}}(d)$ function obtains maximum values randomly over all possible delays, making it not suitable for beamforming. In this case no source delay information can be extracted from the signal and the delays ought to be totally discarded and substituted by others in the surrounding time frames, as will be seen in next section.

C. TDOA Values Selection/Post-Processing

Once the TDOA values of all channels across all meeting have been computed it is desirable to apply a TDOA post-processing to obtain the set of delay values to be applied to each of the signals when performing the weighted-delay&sum as proposed in eq. 1. We implemented two filtering steps, a noisy TDOA detection and elimination (TDOA continuity enhancement), and 1-best TDOA selection from the N-best vector.

1) Noisy TDOA Thresholding: This first proposed filtering step is intended to detect those TDOA values that are not reliable. A TDOA value does not show any useful information when it is computed over a silence (or mainly silence) region or when the SNR of either of the signals being compared is low, making them very dissimilar. The first problem could be addressed by using a speech/non-speech detector prior to any further processing, but prior experimentation indicated that further errors were introduced due to the detector. The selected algorithm applies a simple continuity filter on the TDOA values for each segment c based on their GCC-PHAT values by using a noise threshold Θ_{noise} in the following way:

$$\mathrm{TDOA}^i_n[c] = \begin{cases} \mathrm{TDOA}^i_n[c-1] & \text{if} \quad \text{GCC-PHAT}^i_1[c] < \Theta_{noise} \\ \mathrm{TDOA}^i_n[c] & \text{if} \quad \text{GCC-PHAT}^i_1[c] \geq \Theta_{noise} \end{cases} \tag{5}$$

where Θ_{noise} is defined as the minimum correlation value below which it can be assumed that the correlation is returning feasible results. It is set independently in every meeting as the correlation values are dependent not only on the signal quality but also on the microphone distribution in the different meeting rooms. In order to find an appropriate value for it, the histogram of the distribution of correlation values needs to be evaluated for each meeting. In our implementation a threshold was selected at the value which filters out the lowest 10% of the cross-correlation frames, using the histogram for all cross-correlation values from all microphones in each meeting.

Experimentation showed that the final performance did not decrease when computing a threshold over the distribution of all correlation values together, compared to individual threshold values computed for each channel independently, which would impose a higher computational burden on the system.

2) Dual-Step Viterbi Post-Processing: This second post-processing technique applied to the computed delays is used to select the appropriate delay to be used among the N-best GCC-PHAT values computed previously. The aim here is to maximize speaker continuity avoiding constant delay switching in the case of multiple speakers, and to filter out undesired beam steering towards spurious noises present in the room.

As seen in figure 2 a two-step Viterbi decoding of the N-best TDOA is proposed. The first step consists of a local (single-channel) decoding where the two-best delays are chosen from the N-best delays computed for that channel at every

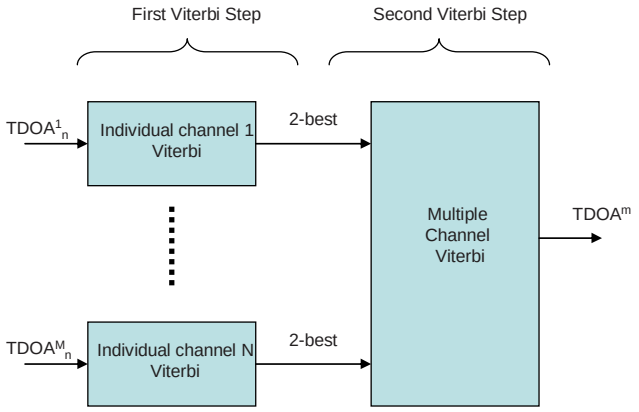


Fig. 2. *weighted-delay&sum double-Viterbi delays selection*

segment. The second decoding step considers all combinations of two-best delays across all channels, and selects the final single TDOA value that is most consistent across all channels. For each step, one needs to define the topology of the state sequence used in the Viterbi decoding and the emission and transition weights to be used. The use of a two-step algorithm is due in part to computational constraints since an exhaustive search over all possible combinations of all N-best values for all channels would easily become computationally prohibitive.

Both steps choose the most probable (and second most probable) sequence of hidden states where each one is related to the TDOA values computed for one segment. In the first step the set of possible states at each segment c is given by the computed N-best values. Each possible state has an emission probability-like value for each processed segment. This value is equal to the $\log$ GCC-PHAT$^i_n[c]$ value for channel i, with $n = 1 \dots N$. No prior scaling or normalization is required as the GCC-PHAT values range from 0 to 1 (given the amplitude normalization performed on the frequency domain in its definition).

The transition weight between two states in step 1 is taken as decreasing linearly with the distance between its delays. Given two nodes, i and j at segments c and $c-1$, respectively, the transition weight for a given channel m is defined as

$$Tr_1^m[i,j;\ c] = \frac{\Delta\text{diff}^m[i,j;\ c] - |\text{TDOA}_i^m[c] - \text{TDOA}_j^m[c-1]|}{\Delta\text{diff}^m[i,j;\ c]} \tag{6}$$

where $\Delta\text{diff}^m[i,j;\ c] = max(|\text{TDOA}_i^m[c] - \text{TDOA}_j^m[c-1]|, \forall i,j)$. This way all transition weights are locally bounded between 0 and 1, assigning a 0 weight to the furthest away delays pair. This implies that only $N-1$ TDOA values will be considered at each segment.

This first Viterbi step aims at finding the two best TDOA values (from the computed N-best) that represent the meeting's speakers at any given time. By doing so it is believed that the system will be able to choose the most appropriate/stable TDOA value for that segment and a secondary delay, which may come from interfering events, e.g. other speakers or the same speaker's echoes. The TDOA values can be any two (not allowing the paths to collapse) of the N-best TDOA values computed previously by the system, and are chosen exclusively based on their distance to surrounding TDOA values and their GCC-PHAT values.

The second pass Viterbi decoding finds the best possible path given the set of hidden states generated by all possible combinations of delays from the two-best delays obtained earlier for each channel. Given a vector $\mathbf{g}(l)$ of dimension $M-1$ (same as the number of channels for which TDOA values are computed) which is the lth combination of possible indexes from the 2-best TDOA values for each channel (obtained in step 1), it is expanded as $\mathbf{g}(l) = [g(l,1) \dots g(l,M-1)]$ where each element $g(l,m) = \{0,1\}$, with 2^{M-1} combinations possible.

One can rewrite GCC-PHAT$^m_{g(l,m)}[c]$, the GCC-PHAT value associated with the $g(l,m)$-best TDOA value for channel m at segment c, which will take values $[0,1]$. Then the emission probability-like values are obtained as the product of the individual GCC-PHAT values of each considered TDOA combination $\mathbf{g}(l)$ at segment c as

$$P_2(\mathbf{g}(l))[c] = \sum_{m=1}^{M} log(\text{GCC-PHAT}^m_{g(l,m)}[c]) \tag{7}$$

which can be considered to be the extension of the individual channel emission probability-like values to the case of multiple TDOA values, where we consider that the different dimensions are independent from each other (interpreted as independence of the TDOA values obtained for each channel at segment c, not their relationship with each other in space along time).

The transition weights are computed in a similar way as in the first step, but in this case they introduce a new dimension to the computation, as now a vector of possible TDOA values needs to be taken into account. As was done with the emission probability-like values, the total distance is considered to be the sum of the individual distances from each element. Assuming $\text{TDOA}^m_{g(l,m)}[c]$ is the TDOA value for the $g(l,m)$-best element in channel m for segment c, the transition weights between two TDOA combinations for all microphones are determined by

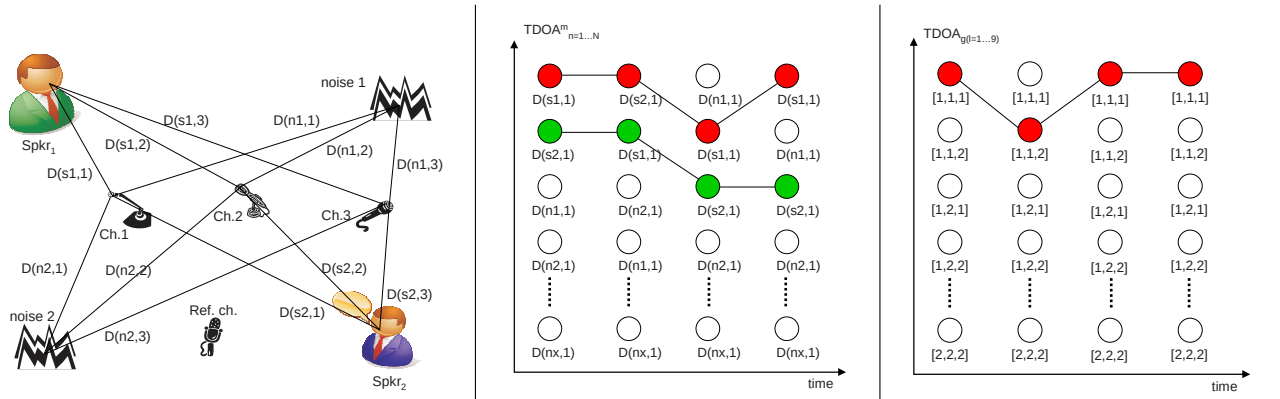
$$Tr_2[i,j;\ c] = \sum_{m=1}^{M} \frac{\Delta\text{diff}[i,j;\ c] - |\text{TDOA}^m_{g(i,m)}[c] - \text{TDOA}^m_{g(j,m)}[c-1]|}{\Delta\text{diff}[i,j;\ c]} \tag{8}$$

where now $\Delta\text{diff}[i,j;\ c] = max(|\text{TDOA}^m_{g(i,n)}[c] - \text{TDOA}^m_{g(j,m)}[c-1]|, \forall i,j,m)$.

This second processing step considers the relationship in space present between all channels, as they are presumably steering to the same position. By performing a decoding over time, it selects the TDOA vector elements according to their distance to nearby vectors.

In both cases, the transition weights are modified (raised to a power) to emphasize thier effect in the decision of the best path. This is similar to the use of word-transition-penalties in an ASR systems. It will be shown in the experiments section that a weight of 25 for both cases appears to optimize the diarization error rate on the development set.

To illustrate how the two-step Viterbi decoding works on the TDOA values, let us consider the example in figure 3a. This



(a) microphones and sound sources layout *(b) Mono-channel 1 first step* *(c) Multi-channel second step*

Fig. 3. *Two-speakers TDOA Viterbi decoding post-processing example*

example shows a situation where four microphones (channels 1–3 and a reference) are used in a room where two speakers ($s1$ and $s2$) are talking to each other, with some overlap speech regions. There is also one or more noisy ($n1, n2, \ldots nx$) events of short duration and room noise in general. Both are represented by a "noise" source. Given one of the microphones as a reference, the delay to each of the other microphones is computed, resulting in delays from speech coming from either speaker ($D(s[1,2],m)$) or from any of the noisy events ($D(nx,m)$) with $m = 1 \ldots 3$.

For a particular segment in the meeting, the N-best TDOA values from the GCC-PHAT cross correlation function are computed. The first Viterbi step determines, for each individual channel, the two-best paths across time for the entire meeting. Figure 3b shows a possible Viterbi trellis for the first step for channel 1, where each column represents the N-best TDOA values computed for one segment. In this example, four segments were considered where the two speakers are overlapping each other, along with some noisy events. For any given segment, the Viterbi algorithm finds the two-best paths (forced not to overlap with each other) according to the distance of the delays to those in the neighboring segments (transition weights) and to their cross-correlation values (emission probability-like values).

In this example, the third computed segment contains a noisy event that is well detected by channel 1 and the reference channel, and therefore it appears as the first in the N-best list. The benefit of using Viterbi decoding is that we avoid selecting this event since its delay differs too much from the best neighboring delays and the fact that both speakers also appear with high correlation. On the other hand, the first and second segments contain the delays for the true speakers in the first and second-best positions, although switched in the segments. This example illustrates a possible case where they cannot be correctly ordered and therefore there is a quick speaker change in the first and second-best delay paths in that segment.

The second step Viterbi decoding adds an extra layer of robustness for the selection of the appropriate delays by considering all the possible delay combinations from all channels. Figure 3c shows the trellis formed by considering, for each segment (in columns), all possible combinations of 2-best delays ($\mathbf{g}(l)[c]$) with dimension 3 (in this example $l = 1 \ldots 9$). For example, the state labeled as [1,2,1] indicates the combination of the 1st-best delay obtained for the first and third microphones, together with the 2nd-best delay on the second microphone.

In this step, only the best path is selected according to the overall combined distances and correlation values among all possible combinations. In this example, the algorithm is capable of solving the order mismatch from the previous step, selecting the delays relative to speaker 1 for all the segments. This is done by maximizing the transition and emission probability-like values between states using Viterbi. In this step, the transition weights are higher for combinations whose TDOA delays are closer in space to each other, i.e. from the same acoustic source, and therefore selecting them ensures steering continuity.

In order to evaluate the correction of the selected TDOA values there are some alternatives, depending on wether we want to make them independent from the signal itself or not. One alternative is to use the resulting signal's SNR. Another alternative is to compute the DER by performing speaker diarization using only the TDOA values.

In conclusion, this newly-introduced two-step Viterbi post-processing technique aims at finding a good tradeoff between reliability (cross-correlation) and stability (distance between contiguous delays). The second of these is perferred since the aim is to obtain an improved signal, avoiding quick changes in the beamforming between acoustic events.

D. Output Signal Generation

Once all information is computed from the input signals, and the optimum TDOA values have been selected, it is time to output the enhanced signal and any accompanying information to be used by the subsequent systems. In this module several algorithms were used to account for the differences between the standard linear microphone array theory and the usual characteristics of meeting room recordings.

1) Automatic Channel Weight Adaptation: In the typical formulation of the weighted-delay&sum processing, the additive noise components on each of the channels are expected to be random processes with very similar power density distributions. This allows the noise on each channel to be statistically cancelled and the relevant signal enhanced 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 one from each other.

In meeting rooms, it is assumed that all of the distant microphones form a microphone array. However, by having different types of microphones there is a change in the characteristics of the signal being recorded and therefore a change in the power density distributions of the resulting additive noises. 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 amplitudes depending on the position of the speaker talking.

This issue is addressed by automatically weighting each channel in the weighted-delay&sum processing in a continuous way during the meeting. This is inspired by the fact that the different channels will have different signal qualities depending on their relative distance to the person speaking, which may change continually during a recording.

The weight for channel m at segment c ($\mathcal{W}_m[c]$) is computed in the following way:

$$\mathcal{W}_m[c] = \begin{cases} \frac{1}{M} & c = 0 \\ (1-\alpha) \cdot \mathcal{W}_m[c-1] + & \\ \alpha \cdot \overline{\text{xcorr}}_m[c] & \text{other} \end{cases} \quad (9)$$

where α is the adaptation ratio, which was empirically set to $\alpha = 0.05$, c is the segment being processed, and $\overline{\text{xcorr}}_m[c]$ is the average of the cross-correlation between channel m and all other channels having all been previously delayed using the selected $\text{TDOA}^m[c]$ value for that channel.

2) Automatic Adaptive Channel Elimination: In some cases, the signal of one of the channels at a particular segment is itself of such low quality that its use in the sum would only degrade the overall quality. This usually happens when the quality of the microphone is poor compared to the others (for example the PDA microphones in the ICSI meeting room recordings as explained in [22]).

In the weighted-delay&sum processing, all available microphones in the room are used and a dynamic selection and elimination of the microphones that could harm the overall signal quality at every particular segment is performed. The previously defined $\overline{\text{xcorr}}_m[c]$ is used to determine the channel quality. If $\overline{\text{xcorr}}_m[c] < \frac{1}{4M}$ then $\mathcal{W}_m[c] = 0$. After checking all the channels for possible elimination, the weights are recomputed so they sum to 1.

3) Channels Sum and Output: Once the output weight has been determined for each channel at a particular segment, all the signals are summed to form the output "enhanced" signal. This output signal needs to be guaranteed acoustic continuity at all times. The theoretical weighted-delay&sum equation as shown in eq. 1, would cause discontinuities in the signal at the segment boundaries due to the mismatch between the signals at the edges.

Therefore, a triangular window is used to smooth and reduce the discontinuity between any two segments, as seen in figure 4. At every segment the triangular filter smooths

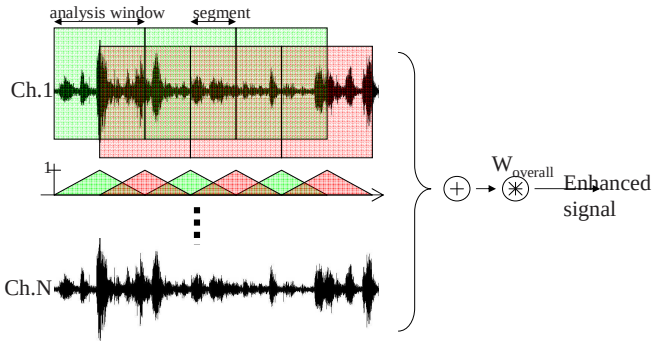


Fig. 4. *Multichannel delayed-signal sum using a triangular window*

the delayed signal using that segment's chosen TDOA value with the signals delayed using the TDOA values from the previous segment. By using the triangular window the system obtains a constant total value without discontinuities. The actual implementation follows equation 10.

$$y[cS+k] = W_{overall}(\alpha[k] \sum_{m=1}^{M} w_m[c] x_m[cS + k - \text{TDOA}^m[c]] + (1-\alpha[k]) \sum_{m=1}^{M} w_m[c] x_m[cS + k - \text{TDOA}^m[c-1]]) \quad (10)$$

where S is the segment sample length, c is the segment being processed and k is the sample within segment being processed.

In the standard implementation, the analysis window overlaps 50% with the segment window as well as the triangular windows used, although it is not necessary for all to use the same overlap values. After all samples from both overlapping windows are summed, the overall weighting factor computed earlier is applied to ensure that the dynamic range of the weighted-delay&summed signal is optimally matched with the available dynamic range of the output file. The resulting enhanced signal is written to a regular PCM file, 16 KHz and 16 bits, which can be processed by any standard speech processing algorithm. In this work it was primarily used for the task of speaker diarization and also some experiments were performed on ASR.

In addition to the acoustic signal, the proposed beamforming system also obtains accurate estimates of the TDOA values for each segment in the meeting. These TDOA values themselves are used to improve speaker diarization performance, as seen in the experiments below.

E. Use of TDOA values for speaker diarization

As explained in [23] and [24], the speaker diarization system in use here is based on an agglomerative clustering technique. It initially splits the data into K clusters (where K must be greater than the number of speakers, and then iteratively merges the clusters (according to the ΔBIC metric described by [25] and later modified by [26]) until a stopping criterion is met. The system uses an ergodic HMM where each state in the HMM is one of the clusters, and each cluster is modeled via a Gaussian Mixture Model (GMM) of varying complexity. Several algorithms are used in order to attempt to obtain the

optimal model complexity and to optimally train each of the models.

When applied to the MDM data, the acoustic features are extracted from the enhanced signal using 19 MFCC coefficients (without any derivatives) and the TDOA values are used without modification.

In order to use the TDOA values to improve diarization, we use a separate set of GMMs to model the TDOA features. The acoustic and TDOA streams share the same speaker clustering information, but each set of GMMs are trained on the data coming from the two separate streams. The combination of both contributions is done at the likelihood level and used in the Viterbi decoding and in the BIC computation steps as a weighted sum of each of the individual log likelihood values. The relative stream weights are obtained automatically by using an adaptive algorithm based on the BIC values as described in [27].

III. EXPERIMENTS

The acoustic beamforming system presented in this work was created for use in the speaker diarization task for the meetings environment. In this section we present experiments to show the usefulness of the techniques introduced in this work with respect to the speaker diarization task, and we also present some comparative results for a speech recognition task.

All databases used in these experiments come from the NIST RT evaluations from 2004–2006: 34 meeting excerpts in total. In all cases only the conference room domain data was used (where a conference is considered to be a meeting where multiple participants interact around a meeting table). Data from RT2004 and RT2005 were used for development (22 multichannel meetings + 4 monochannel meetings), and data from RT2006 was used for testing (8 multichannel meetings). The excerpts were recorded in various physical layouts, using different types of microphones, and included data from ICSI, NIST, LDC, CMU and others.

The evaluation metrics used are the standard NIST Diarization Error Rate (DER) for speaker diarization and Word Error Rate (WER) for speech recognition. The DER is the percentage of time that the system misassigns speech (either between speakers or speech/non-speech) and includes the regions where two (or more) speakers speak simultaneously causing overlapping speech. The WER is the percentage of erroneous words in a transcript. In the case of speech recognition, the reference transcriptions were created manually using the signals recorded from each of the meeting participant's headset microphone. For the diarization experiments, forced alignments were computed using these transcriptions in order to obtain the reference speaker information. All DER results are computed taking into account errors in the overlapping speech regions.

A. Speaker Diarization Experiments

We performed two sets of experiments. First, the acoustic beamforming algorithms were tested using only the multichannel meetings. The baseline system for the experiments presented in this first set of experiments is the full system as used in the RT06s evaluation [23] (this includes the beamforming, as explained here, a speech/non-speech module and a single channel speaker diarization module.) Using this baseline system, we then modify just the key beamforming algorithms presented in this work to show their effect in isolation. These tests only take into account the acoustic data output from the beamforming (i.e. no TDOA values were used.)

A second set of experiments uses all of RT meetings available (both single channel and multichannel), an improved speaker diarization module and a speech/non-speech module for each signal to show how the acoustic beamforming improves results compared to using the most centrally located microphone in the meeting (defined by NIST as the SDM channel.) These diarization tests use both the acoustic data and the TDOA values to improve diarization as shown in [28].

Tables I and II summarize the results for the first set of tests, comparing a full beamforming system (labelled RT06s baseline) with a system where some of the proposed algorithms have been removed, while keeping all other modules constant. For each system, the DER is computed for the development and evaluation sets as well as the absolute DER percentage variation and the percentage variation versus the baseline system ($\Delta\%$). So, a negative value indicates an improvement over the baseline system, which means that the use of the technique lowered the performance of the baseline system (i.e. not using the technique may represent a potential improvement.). On the test results in Table II the last column shows the measured significance parameter Z for each system compared to the baseline. Such test is essentially a t-test which applies the Matched Pairs Sentence-Segment Word Error (MAPSSWE) test introduced by [29] and implemented by NIST in [30]. In Diarization we defined each segment to have 0.5 seconds. For a significance level of 5% the differences are considered significant when $Z > 1.96$

F&S system	Development		
	DER	ΔDER	Δ%
RT06s baseline system	17.15%	-	-
1)Hand-picked ref. channel	17.09%	−0.06	−0.3%
2)No noise threshold	18.31%	1.15	6.3%
No TDOA-Viterbi algorithm	17.16%	0.01	0%
No TDOA post-processing	18.72%	1.57	8.3%
3)No adaptive weights	17.48%	0.33	1.8%
No channels elimination	17.14%	−0.01	0%

TABLE I
Diarization Error Rate (DER) comparison on the development set for each of the proposed algorithms

F&S system	Evaluation			
	DER	ΔDER	Δ%	Z
RT06s baseline system	22.92%	-	-	-
1)Hand-picked ref. channel	22.49%	−0.43	−1.9%	0.87
2)No noise threshold	23.38%	0.46	2.0%	0.85
No TDOA-Viterbi algorithm	23.63%	0.71	3.0%	2.57
No TDOA post-processing	22.25%	−0.67	−2.9%	1.98
3)No adaptive weights	24.06%	1.14	4.7%	4.25
No channels elimination	23.91%	0.99	4.1%	2.02

TABLE II
Diarization Error Rate (DER) comparison on the evaluation set for each of the proposed algorithms

1) Meeting Information Extraction Tests: The first comparison corresponds to the selection of the reference channel to use in the TDOA calculation by taking into account prior

information- using the SDM channels as defined by NIST for each excerpt. By using automatic selection of the reference channel (as is done in the baseline system) the results are slightly worse. Although the DER of the development set is almost equal to the hand-picked reference channel, the performance on the eval set shows a relative improvement in DER of 1.87%. We consider that it is still preferable and more robust to use the automatic selection of the reference channel, as it then becomes possible to use this system in areas other than the RT evaluation data, where there might not be any prior information on which microphone to select as the reference. Furthermore, on the test set the significance test shows that such difference between systems is not significant enough ($Z < 1.96$).

2) TDOA Values Selection Tests: The following three systems correspond to algorithms in the post-processing module, which includes the noise thresholding and the TDOA values stability algorithms. When comparing the full RT06s baseline system with a system that doesn't use any of the postprocessing algorithms, we obtained mixed results depending on the data set. For the development set, the postprocessing algorithms improve results by 8.3% relative, while on the evaluation set, performance is 2.9% worse. In order to fully study such differences, we examine the effects of not using either the noise thresholding algorithm or the the TDOA continuity algorithm.

On the one hand, the noise thresholding algorithm acts as a simple speech/non-speech detector at the beamforming level. Initial tests were performed to try using a more sophisticated detector, but in the end, it was not used as the scores were about 10% worse, and it just complicated the system. When studying the effect of not using noise thresholding, we observed that in both the development (6.3% relative) and the evaluation sets (2.0% relative), there was a gain in performance. The noise threshold percentage was initially set to 10% (without performing any optimization experiments), which accounted for all outliers which we wanted to eliminate. For some cases, a higher value of 20% did give slightly better performance and values lower than 10% did not show as much improvement.

The final implementation of the noise threshold takes into account the histogram of the GCC-PHAT values on the current meeting, rather than setting a fixed threshold as reported in [8]. This is done to attempt to compensate for noisy meetings (like some LDC recordings in the NIST RT datasets), where the best threshold is not the same as in the less noisy recordings.

On the other hand, the TDOA continuity algorithm is compared to not performing any continuity processing. The use of this algorithm did not improve performance on the development set, but showed a 3% relative improvement for the test set. In order to process the double Viterbi TDOA decoding, the internal weight variables were both set to 25 in the RT06s baseline system. Further testing performed after the RT06s evaluation showed that setting the first weight to 15 improved the development set results, but worsened the evaluation results. A homogeneous value of 25 seems to be a safe selection for both datasets. For a more complete study of the variation of this parameter refer to [24].

Another parameter that needs adjusting in the continuity algorithm, is the number of N-best values to be considered by the algorithm when selecting the optimum TDOA value. The first Viterbi step does a local selection within each channel from the N-best possible TDOA values to the 2-best, which then are considered by the second Viterbi in a global decoding using all of the TDOA values from all channels. The number of possible initial TDOA values is a parameter that describes how many possible peaks in the GCC-PHAT function have to be considered by the first Viterbi. The selection of the optimal number of initial N-best values needs to account for concurrent acoustic events while avoiding false peaks in the GCC-PHAT function. The default value for N was set to 4 in the RT06s system based on tests performed on development data and based on the DER and a signal to noise SNR measure.

Overall, the two individual TDOA selection algorithms each improve performance independently. For the development set, the combinantion of the two techniques shows an improvement that is larger than the sum of individual improvements. On the evaluation set each individual algorithm performs well in isolation while the combined performance is worse. The significance test of this system compared to the baseline is passed on both development and test cases ($Z = 1.98$ in development, $Z = 2.31$ in test). A per-meeting basis analysis should be performed in order to assess particular cases where these algorithms do not perform well together.

3) Output Signal Generation Tests: The final two results show tests performed when not using the algorithms related to output signal generation. When no channel weights are used, a $\frac{1}{N}$ constant weight is applied to all channels. The DER improves by 1.8% relative by using the relative channel weights on the development set and by 4.7% relative on the evaluation set. So it appears that this algorithm is beneficial to the system and it does not impose a significant computational burden on the system.

Experiments with eliminating frames of data from the "bad" channels show that the DER does not change for the development set, but improves by 4.1% relative for the evaluation set. We believe this is due to the dependency of this algorithm on the relative quality of microphones in each recording setup. When all microphones are of a similar quality, none of them looses frames and therefore the results should be the to the system where the algorithm was not used. Both algorithms passed the significance test on the test data.

4) Overall Acoustic Beamforming Tests: Now that we have examined the usefulness of each of the individual algorithms involved in the generation of an enhanced output signal, we will attempt to assess how well the beamforming system can take advantage of the multiplicity of recording channels in a meeting environment. To do this, we will use both the MFCC and TDOA values, comparing the output of the speaker diarization system for the Multiple Distant Microphone condition (MDM+TDOA), with that of the most centrally located Single Distant Microphone (SDM, as defined by NIST) condition.

In figure 5 we see improvements of 41.15% and 25.45% relative on the development and test sets respectively. This is due to several factors: the improved quality of the beam-

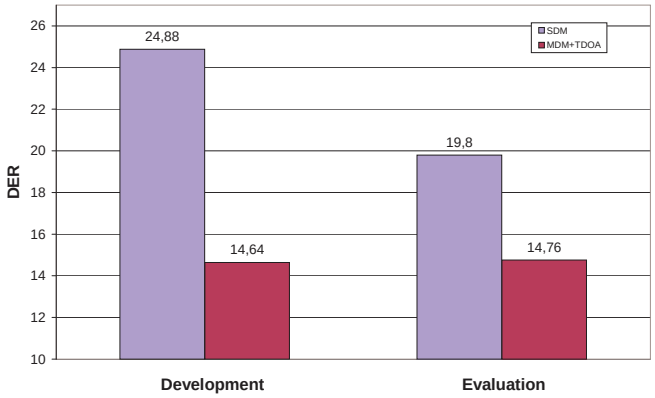


Fig. 5. *Using a single microphone versus multiple microphones combined via acoustic beamforming on meetings*

formed signal, which also propagates to the speech/non-speech module, and the use of the TDOA values from the beamforming adds additional information about the current speaker's position in the room. Both these issues are the result from applying the beamforming algorithm presented in this paper to the diarization task: the enhanced acoustic signal and the information about speakers positions brought by the TDOA values (which is otherwise lost when collapsing all channels acoustic data into one).

In order to isolate the improvements resulting from using the enhanced acoustic signal and from inserting the TDOA values in the speaker diarization module, we will be applying the system used to compute figure 5 to the development set. Using only the acoustic features for diarization (no TDOA values) we obtained a 19.04% DER. This shows a similar, incremental improvement coming from the beamformed signal (first) and adding to it the information from the TDOA in diarization (second).

On other data sets, we observe different behavior. Namely, adding TDOA values results in a much larger improvement than just using the MFCC features computed from the beamformed signal alone. This is due to the heterogeneity of the acoustic channels which are to be beamformed. In some meeting setups, although TDOA values can be well estimated, the signal quality of some channels can degrade the overall acoustic output. Adaptive weighting and channel elimination algorithms help to obtain always an output signal which is of more quality than any of the individual ones, although in some cases this improvement might be minimal. For more comparisons and experiments refer to [24].

In order to study the significance of these results we apply the test described before on the test data. We obtain a significance factor $Z = 8.26$ comparing the SDM system with the MDM+TDOA system and a $Z = 8.1$ comparing the MDM with the MDM+TDOA systems, indicating both are very significant results and not due to randomicities.

B. Speech Recognition Experiments

The beamforming system developed for the speaker diarization task was also used to obtain an enhanced signal for the ASR systems that ICSI and SRI presented at the RT NIST evaluations. For RT05s the same beamforming system was used for ASR and for speaker diarization. As explained in [31], evaluating on the RT04s eval set, and excluding the CMU mono-channel meetings, the new beamforming outperformed the previous version of the ICSI beamforming system by 2.8% absolute (from 42.9% word error rate to 40.1%). The previous beamforming system in use at ICSI was based on delay&sum of full speech segments (obtained from a speaker segmentation algorithm).

For the RT06s system the beamforming module was tuned separately from the diarization module to optimize for Word Error Rate (WER), leading to a system which was more robust than the RT05s beamforming system. Although acoustic beamforming attempts to optimize the enhanced signal's SNR, the use of the enhanced signal in these two systems behaves slightly differently because the two systems are evaluated using different metrics, one based on time alignment and the other on word accuracy. In fact, in [24] it is shown that SNR and DER behave differently and therefore optimizing the beamforming system with one metric doesn't necessarily improve performance on the other metric. In fact, separate tuning was not found to be crucial as only about 2% relative improvement on WER was gained compared to using a common beamforming system.

dataset	SDM	MDM	ADM	MM3A
RT05s	47.7%	45.8%	38.6%	–
RT06s	57.3%	55.5%	51%	56%

TABLE III
WER using the RT06s ASR system including the beamformer

As seen in [32], and reproduced in table III, the RT05s and RT06s datasets were used to evaluate the RT06s ASR system. In both datasets, there is an improvement of almost 2% absolute improvement over SDM by using beamforming in the MDM condition. These two ASR systems are identical except that the system for MDM uses the weighted-delay&sum algorithm, along with some minor tuning parameters which were optimized for each condition.

This improvement becomes much larger between the MDM and ADM cases, where the improvement is exclusively due to the fact that the acoustic beamforming was performed using many more microphones (in the ADM case).

The multiple mark III microphone arrays (MM3a) were available for the RT06s evaluation data on lecture rooms. Tests performed comparing results with other state of the art beamforming systems showed that the proposed beamformer performed very well.

IV. Conclusion

When performing speaker diarization on recording from the meetings domain, we often have recordings available from multiple microphones. There have been several approaches in recent years trying to take advantage of this information. However, these approaches have had only limited success compared to using only a single, most centrally located, microphone. In this paper we present an approach, based on popular acoustic beamforming techniques, to obtain a single enhanced signal and speaker-position information from a number of microphones. We have proposed several novel algorithms to obtain improved signal quality, under most conditions, for

the task of speaker diarization. Additionally, we have shown improvements due to the use of between-channel delay values as a form of spacial information for the diarization task. Tests performed on NIST rich transcription data showed a significant reduction in error for the diarization task compared to using just a single microphone. In addition, tests using the same beamforming system in a speech recognition task also showed improvements over previous beamforming implementations. We believe that the proposed use of acoustic beamforming for speaker diarization is an important step towards the goal of filling the performance gap between meetings data and broadcast news data in the task of speaker diarization.

ACKNOWLEDGMENT

This work was made possible thanks to the AMI training program and the Spanish visitors program, both of which allowed X. Anguera to visit ICSI for two years. Special thanks to to the people responsible for these initiatives. Thanks also to Marc Ferras for various technical help during the development of these algorithms and to Jose M. Pardo for his contribution in the use of delays in speaker diarization.

REFERENCES

[1] S. Cassidy, "The Macquarie speaker diarization system for RT04s," in *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada, 2004.

[2] D. van Leeuwen, *The TNO Speaker Diarization System for NIST RT05s for Meeting Data*, ser. Machine Learning for Multimodal Interaction (MLMI 2005) in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 3869, pp. 440–449.

[3] D. van Leeuwen and Marijn Huijbregts, *The AMI speaker diarization system for NIST RT06s meeting data*, ser. Machine Learning for Multimodal Interaction (MLMI 2006) in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 4299, pp. 371–384.

[4] 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.

[5] 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.

[6] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J.-F. Bonastre, *NIST RT05s evaluation: Pre-Processing Techniques and Speaker Diarization on Multiple Microphone Meetings*, ser. Machine Learning for Multimodal Interaction (MLMI 2005) in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 3869, pp. 428–439.

[7] C. Fredouille and G. Senay, *Technical improvements of the E-HMM based speaker diarization system for meetings records*, ser. Machine Learning for Multimodal Interaction (MLMI 2006) in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 4299, pp. 359–370.

[8] X. Anguera, C. Wooters, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. ASRU*, Puerto Rico, USA, November 2005.

[9] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, *Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System*, ser. Machine Learning for Multimodal Interaction (MLMI 2005) in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 3869, pp. 402–414.

[10] B. van Veen and K.M. Buckley, "Beamforming: A versatile approach to spacial filtering," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1988.

[11] H. Krim and M. Viberg, "Two decades of array signal processing research," *IEEE Signal Processing Magazine*, pp. 67–94, July 1996.

[12] J. G. Fiscus, J. Ajot, M. Michet, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *NIST 2006 Spring Rich Transcrition Evaluation Workshop*, Washington DC, USA, May 2006.

[13] "Beamformit: Open source acoustic beamforming software," http://www.icsi.berkeley.edu/~xanguera/beamformit, 2007.

[14] 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.

[15] D. Johnson and D. Dudgeon, *Array signal processing*. Prentice Hall, 1993.

[16] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.

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

[18] Wiener and Norbert, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, 1949.

[19] 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 *Proc. ICSLP*, Jeju Island, Korea, October 2004.

[20] "NIST rich transcription evaluations," http://www.nist.gov/speech/tests/rt, 2006.

[21] "ICSI meeting recorder project: Channel skew in ICSI-recorded meetings," http://www.icsi.berkeley.edu/~dpwe/research/mtgrcdr/chanskew.html, 2006.

[22] 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," in *ICCASP*, Montreal, 2004.

[23] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in *Proc. ICSLP*, Pittsburgh, USA, September 2006.

[24] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Universitat Politecnica de Catalunya, December 2006.

[25] S. S. Chen and P. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proc. ICASSP*, vol. 2, Seattle, USA, 1998, pp. 645–648.

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

[27] X. Anguera, C. Wooters, and J. Hernando, "Automatic weighting for the combination of TDOA and acoustic features in speaker diarization for meetings," in *Proc. ICASSP*, April 2007.

[28] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Proc. ICSLP*, September 2006.

[29] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.

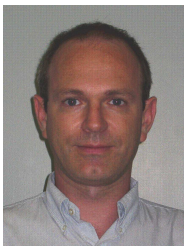
[30] D. Pallett, W. Fisher, and J. Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Proc. ICASSP*, vol. 1, 1990, pp. 97–100.

[31] 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*, ser. Machine Learning for Multimodal Interaction (MLMI 2005) in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 3869, pp. 463–475.

[32] A. Janin, A. Stolcke, X. Anguera, K. Boakye, O. Cetin, J. Frankel, and J. Zheng, *The ICSI-SRI Spring 2006 Meeting Recognition System*, ser. Machine Learning for Multimodal Interaction (MLMI 2005) in Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2006, vol. 3869, pp. 444–456.

Xavier Anguera Xavier Anguera Miro (Ing. [MS]. 2001 UPC University, Dr. [PhD] 2006 UPC University, with a thesis titled "Robust Speaker Diarization for Meetings". After graduating from his Masters he worked until 2003 for Panasonic Speech Technology Laboratory in Santa Barbara California, where he developed a Spanish TTS system and did research on speaker recognition. At the end of 2003 Dr. Anguera started a PhD program in the Department of Signal Theory and Communications of the UPC, working on speaker diarization. From September 2004 until September 2006 he was visiting ICSI where he worked on speaker diarization for meetings and participated in several NIST RT evaluations. He is currently with Telefnica I+D pursuing research on speaker technologies and actively participating in Spanish and European research projects. His interests cover the areas of speaker technology and automatic indexing of acoustic data.

Chuck Wooters Charles Wooters holds a BA and MA in Linguistics from the University of California at Berkeley. He received his PhD from Berkeley in "Speech Recognition" in November 1993. This interdisciplinary program spanned the departments of Computer Science, Linguistics and Psychology. After graduating from Berkeley, he went to work for the U.S. Department of Defense (DoD) as a speech recognition researcher. In April 1995, he joined the software development group at Computer Motion Inc. in Goleta California. While at Computer Motion, Dr. Wooters developed the speech recognition software systems that were used in Aesop and Hermes. In April 1997, Dr. Wooters returned to the DoD where he continued to perform research in large vocabulary continuous speech recognition. In 1999, he joined the Speech and Natural Language group at BBN where he led a small group of researchers working on government-sponsored research in speech and natural language processing. In 2000, Dr. Wooters joined the speech group at the International Computer Science Institute in Berkeley where he continues to perform research, specializing in automatic speaker diarization.



Javier Hernando Javier Hernando, M.S. and Ph.D. degrees in Telecommunications Engineering from the Technical University of Catalonia (UPC), Spain, in 1988 and 1993, respectively. Since 1988 he is with the Department of Signal Theory and Communications of the UPC, where he is now an Associate Professor and member of TALP ((Research Centre for Language and Speech). He was a visiting researcher at the Panasonic Speech Technology Laboratory, Santa Barbara CA, during the academic year 2002-2003. His research interests include robust speech analysis, speech recognition, speaker verification and localization, oral dialogue and multimodal interfaces. He has about 150 publications in book chapters, review articles and conference papers on these topics. Dr. Hernando received the 1993 Extraordinary Ph.D. Award of UPC. He has led the UPC team in several European, Spanish and Catalan projects.