Speaker Diarization for Multiple-Distant-Microphone Meetings Using Several Sources of Information

José M. Pardo, *Senior Member, IEEE*, Xavier Anguera, *Member, IEEE*, and Charles Wooters, *Member, IEEE*

Abstract—Human-machine interaction in meetings requires the localization and identification of the speakers interacting with the system, as well as the recognition of the words spoken. A seminal step toward this goal is the field of rich transcription research, which includes speaker diarization together with the annotation of sentence boundaries and the elimination of speaker disfluencies. The subarea of speaker diarization attempts to identify the number of participants in a meeting and create a list of speech time intervals for each such participant. In this paper, we analyze the correlation between signals coming from multiple microphones and propose an improved method for carrying out speaker diarization for meetings with multiple distant microphones. The proposed algorithm makes use of acoustic information and information from the delays between signals coming from the different sources. Using this procedure, we were able to achieve state-of-the-art performance in the NIST spring 2006 rich transcription evaluation, improving the Diarization Error Rate (DER) by 15 percent to 28 percent relative to previous systems.

Index Terms—Speech source separation, speaker diarization, speaker segmentation, meetings recognition, rich transcription.

1 Introduction

Human-machine interaction in meetings requires the localization and identification of the speakers interacting with the system, as well as the recognition of the words spoken. A seminal step toward this goal is the field of rich transcription research, which includes speaker diarization together with the annotation of sentence boundaries and the elimination of speaker disfluencies. The rich transcription research area was initially motivated by the problem of speech transcription for increasingly complex audio sources: telephone conversations, broadcast news (BN), and meeting domains. The goal is to annotate the data with as much detail as possible with regard to speaker turns, sentence units, and so forth, for possible downstream applications (for example, indexing, translation, and so forth). With this ambitious objective in mind, several years ago, the US National Institute of Standards and Technology (NIST) started a series of evaluations to tackle this problem and defined the field of rich transcription to complement speech-to-text transcription or speech recognition [1]. One of the tasks defined by NIST was speaker diarization. For the meeting domain, speaker diarization is the task of identifying the number of participants in the meeting and creating a list of speech time intervals for each participant. It is important to note that diarization as defined by NIST is carried out without any prior knowledge of the location or identity of the speakers in the room, the location or quality of the microphones, or the details of the acoustics of the room. Although prior knowledge of the microphone locations would permit precise speaker localization and segmentation, as used in [2], [3], [4], and [5], these types of methods cannot be used in this task since microphone locations are not available in the NIST diarization evaluation.

One use of speaker diarization is to aid the transcription task. Instead of just transcribing a recording into unorganized text, the transcription is annotated with a different label for each speaker. Later, if we knew in advance the set of possible speakers that would appear in the recording, we could use a speaker verification algorithm to assign an identified speaker to every label. A transcription annotated in this manner is more readable and useful. It could also be used for automatic speaker indexing of audio documents.

A second possible application of speaker diarization is to aid in the application of adaptation techniques to speech recognition. Once we know the regions that correspond to a speaker, we can adapt the recognizer to do a better job by using speaker-dependent speech recognition. A good introduction to the topic of audio diarization and speaker diarization is given in [6].

Finally, the methods used for carrying out speaker diarization—particularly, if they use delays between signals—will surely be relevant in the more difficult task of online speaker diarization (identifying who is speaking in a meeting and what is being said), particularly if only multiple microphones are available (with no video cameras)

- *J.M. Pardo is with the ETSI de Telecomunicación, Universidad Politecnica de Madrid, Ciudad Universitaria, 28040 Madrid, Spain. E-mail: pardo@die.upm.es.*
- *X. Anguera is with Telefónica I+D, Via Augusta 177, 08021 Barcelona, Spain. E-mail: xanguera@tid.es.*
- *C. Wooters is with the International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, CA 94704. E-mail: wooters@icsi.berkeley.edu.*

Manuscript received 12 Sept. 2006; revised 6 Mar. 2007; accepted 16 Apr. 2007; published online 10 May 2007.
Recommended for acceptance by R.C. Guido, L. Deng, and S. Makino.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TCSI-0367-0906.

0018-9340/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

and several speakers can talk at the same time (there is overlapping speech).

Since 2002, NIST has included speaker diarization as one of the tasks evaluated in the context of rich transcription of meetings [7], which evolved from speaker diarization for BN and telephone conversations. In 2002, the evaluation was carried out using a single distant microphone (SDM). Since 2004, the primary condition in the evaluation has moved to multiple distant microphones (MDM).[1]

In the rest of this section, we present an introduction to the methods used in speaker diarization, classifying the topic into two different sections. Section 1.1 explains diarization methods for tasks in which a single (distant) microphone is available. Section 1.2 tackles the topic of speaker diarization when several distant microphones can be used. Finally, a summary of the contents of the rest of the paper is given.

1.1 Speaker Diarization for Meetings Using an SDM

In general, the approaches used in the literature for speaker diarization using an SDM or a single recording signal have their basis in previous audio segmentation and diarization systems from the BN domain [6], [8], [9], [10], [11]. A good overview of this topic has recently been published [11]. The process usually starts by finding and eliminating nonspeech frames from the recording. This task is sometimes difficult since nonspeech may include silence, music, laughter, breath, lip smack, paper shuffling, and so forth. There are several ways of accomplishing this. The first is to use maximum likelihood classification with two Gaussian Mixture Models (GMM): one for speech and one for silence and other sounds, as in [12]. Other authors explicitly model noise and music [13], [14]. Finally, speech detection can also be made using a phone recognizer or a word recognizer [15].

The next step in the process is to find points of acoustic changes in the signal and create acoustically homogeneous segments (segmentation). This is done by analyzing adjacent windows of data and calculating the distance between them [17]. In integrated systems, such as the one used at our laboratory at the International Computer Science Institute (ICSI), there is no need to carry out this step explicitly [18]. The next step, which is especially relevant in BN, although it is optional, is to classify these segments into male/female, narrow band/high band, and so forth. The homogeneous segments are then hierarchically clustered to combine acoustically similar segments appearing at different times in the show. One limitation of this method is that errors made in the segmentation step cannot be corrected later. More advanced systems resegment the signal after the clustering and further cluster the segments in an iterative process [10].

The Laboratoire d'Informatique d'Avignon (LIA) system and the Communication Langagiere et Interaction Personne Systeme-Institut d'Informatique et Mathematiques Appliquées de Grenoble (CLIPS) system [10], [16] are good examples for the comparison of both approaches: step-by-step versus integrated. The CLIPS system is a sequential system based on speaker change detection followed by hierarchical clustering. It uses the global likelihood ratio (GLR) for acoustic change detection and clustering distance and the Bayesian Information Criterion (BIC) [19] as a stopping criterion. It also uses maximum a posteriori (MAP) adaptation to train the cluster models from a background model. The LIA system, in contrast, is an integrated approach that uses a dynamic Hidden Markov Model (HMM) to generate speaker clusters top down, retraining the models and resegmenting the show every time a new speaker is added. Run separately, the LIA system outperforms the CLIPS system, 16.9 percent to 19.3 percent diarization error rate (DER) on the RT03s data set [20]. The two laboratories created a joint system using the CLIPS system as the first module followed by the LIA model, obtaining a 12.9 percent error and winning the RT03s evaluation.

Tranter and Reynolds present two systems in [8]: one from the Cambridge University Engineering Department (CUED) and one from the Massachusetts Institute of Technology-Lincoln Labs (MIT-LL). The CUED system uses a step-by-step approach, although segmentation is not carried out with an acoustic change detector but with a phone recognizer. Clustering is carried out using arithmetic harmonic sphericity as the distance metric and they compare three different stopping criteria with the winner being BIC. MIT-LL also uses a step-by-step method, where the acoustic change detector system is based on adjacent window comparisons using BIC and the clustering and stopping criterion also uses BIC. The paper also proposed a "plug-and-play combination" using the components of both systems with the best DER combination using CUED for segmentation and MIT-LL for clustering.

One of the best systems recently published for BN was presented by the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) in [9]. Their system is an integrated system which uses innovative methods to improve the performance. They include a speaker identification module to carry out cluster adaptation and a speech recognition module to refine the final segment labels. The system uses all possible information and data available to do the task, as well as training data from the same domain. This system was the winner of the RT04f evaluation [21].

The meeting domain differs from BN as the topics are highly diverse, the participants have idiosyncratic relationships and vocabularies, the meetings are highly interactive, and there can be simultaneous speech from multiple speakers. Furthermore, distant microphones are susceptible to reverberation and background noise. Consequently, the problem is much more difficult than in the BN domain, although, in BN, the number of speakers may be much higher. In 2002, NIST conducted an evaluation of speaker diarization in the meeting domain under the SDM condition. Although tests carried out since 2002 have considered MDM as the primary condition, the methods applied to SDM or previously to BN may be considered a first step toward the development of algorithms for MDM.

[1] For NIST Meetings, a distant microphone is a microphone located in on a table or in a wall at a minimum distance from the mouth of any speaker (several centimeters) as opposed to an individual head mounted microphone (IHM) where the microphone is close to the mouth of the speaker.

There has been extensive research at ICSI in the last few years in the area of meeting recognition, including speech recognition and speaker diarization [17], [18], [12], [22], [23], [24], [25]. The basic method used at ICSI for SDM can be considered an integrated approach, which models the utterance using an ergodic HMM with a number of states equal to the initial number of speaker clusters (K). Each state in the HMM contains a sequence of substates that imposes a minimum duration on the cluster. Within a state, each of the substates uses a probability density function (PDF) modeled by a GMM with a diagonal covariance matrix [18], [12]. Essentially, the process consists of two modules: an initialization and an integrated segmentation and clustering module. The initialization requires an informed guess at the maximum number of speakers (K) that are likely to occur in the data. The data is then divided into K equal-length segments and each segment is assigned one GMM. Each GMM is then trained using its associated data. These models are then used to seed the following clustering module, which uses agglomerative clustering and segmentation steps in an iterative loop.

1.2 Speaker Diarization for Meetings with MDM

The task of speaker diarization for meetings with MDM should be easier compared to an SDM because 1) there are redundant signals—one for each microphone—that can be used to enhance the signal, even if some of the channels have a very poor signal to noise ratio (SNR), and 2) the signals contain information on the spatial position of the audio source (speaker). In a previous work [26], a processing technique using the time delay of arrival (TDOA) was applied to the different microphone channels by delaying and summing the channels to create an enhanced signal. With this enhanced signal, the DER was improved by 3.3 percent relative compared to the SDM error for the RT05s evaluation set, 23 percent relative for the RT04s development set, and 2.3 percent relative for the RT04s evaluation set. (See [7] for more information on both the data sets and the task.)

The use of speaker location information to carry out speaker diarization can be divided into two main categories, one that uses microphone locations and microphone geometry to establish explicit speaker location and one that cannot use microphone location information (because it is not available) as in the task that we are addressing. The benefits of using explicit speaker locations are that a precise speaker segmentation can be carried out and that the tracking of moving speakers is also feasible. The disadvantage is that exact microphone location and the synchronization of the signals are also needed. If microphone locations are not available, the task of speaker diarization is more difficult, as well as more generic. In the first category, we can cite the work done by Lathoud et al. [4], [3]. In [3], an algorithm is proposed that can track multiple moving speakers using event location cues alone. The algorithm is particularly efficient in tracking concurrent events such as speakers' overlaps (one of the most difficult problems to tackle) in real time. The algorithm does not need any prior knowledge of the speakers' location, as in [4]. In the second category, the only work that we are aware of is that of Ellis and Liu in [27]. They used the cross-correlation between channels to find a peak that corresponds to the time delay between two channels and they then clustered the time delays to create homogeneous segments of frames. The result they reported for the RT04s development set was a 62.3 percent DER^*.[2]

When several microphones are available, it is obvious to try to merge acoustic information from the speakers, as well as speakers' location. TDOA features permit short-term speaker segmentation but do not provide any speaker identity information. On the other hand, acoustic features provide long-term speaker identity but require minimum durations to build reliable acoustic models. Again, in this part, we have to consider whether microphone locations are available or not as two different tasks. Ajmera et al. [2] and Lathoud [5] use microphone location information and combine both features. They demonstrate that the fusion of these two types of information improves speaker diarization. For the NIST task where microphone locations are not available, a different approach is needed and we are not aware of any published work that addresses the combination of acoustic features and TDOA features.

In the first part of this paper, we present several experiments to determine to what extent the TDOAs by themselves can be used to segment and cluster the different speakers in a room. We have tried to develop a system that is robust to the changes in the conditions of the meeting, room, microphones, speakers, and so forth, although our method assumes that the speakers do not move far from their seat. We present a method that only uses the delays to segment and cluster the speakers. In our method, we obtain a diarization error (DER^*) [7] of 35.73 percent on the same set of meetings used by Elis and Liu, a 42.64 percent relative improvement. We also provide the results obtained from other data sets, including RT05s and RT06s.

In the second part of the paper, we present an original method to combine the acoustic front-end features, MFCC, with the TDOA features to obtain an enhanced segmentation. By merging both TDOA and acoustic features, we have been able to improve the baseline results (using only acoustic features) by 16.34 percent relative on the RT05s evaluation set, 27.52 percent relative on the Devel06s development set (see the explanation of the data sets below), and 15.10 percent relative on the RT06s evaluation set. This method was a top performer in the RT06s evaluation [28]. For information on the databases mentioned in this section, see [29].

The paper is organized as follows: In Section 2, we describe the basics of our system. In Section 3, we present the data set used and the evaluation metric. Section 4 explains the basic diarization system using only acoustic data and multiple microphones. In Section 5, we introduce our novel method to use interchannel differences exclusively to carry out speaker diarization, which shows good performance compared to previously published results on the same data and task. This is the first contribution of this paper. Section 6 explains the second novelty, the technique we used to combine acoustic information and delay information to improve the performance of the ICSI system.

[2] The equivalent that they used is DER minus false alarm (in NIST terminology); we have called it DER^*

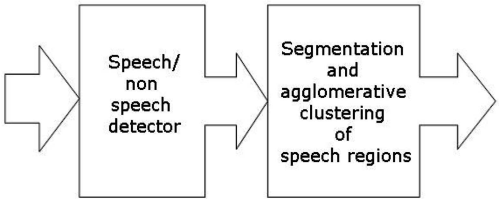

Fig. 1. Speaker diarization system architecture.

In Section 7, we discuss the results obtained and the advantages and drawbacks of our proposal. Section 8 describes the conclusion.

2 System Description

2.1 System Architecture

The general system architecture is shown in Fig. 1. First, the utterance is segmented into speech and nonspeech regions. The speech is segmented into homogeneous chunks and then clustered and resegmented iteratively until a stopping criterion is reached.

2.2 Speech/Nonspeech (SNS) Detector

One of the most important tasks in the process of speaker diarization is the separation of speech from all of the other audio components, including silence, background noise, and nonspeech sounds such as laughs, coughs, breaths, and so forth. In fact, the task is so important that NIST has decided to evaluate it separately from speaker diarization in a task known as Speech Activity Detection (SAD). The detection of speech plays a crucial role in both speaker diarization and speech recognition because errors made at this stage cannot be recovered later. Because of the method used to calculate DER, every speech detection error is propagated to the end of the process either as a false alarm (FA) error—speech is detected where there is no true speech—or a missed speech error—nonspeech is detected where there is speech in the reference.[3]

The question that arises is: What is the best method to accomplish this task? Clearly, the approach is dependent on the application. It is also important to know whether there is training data available.

Two methods have been used at ICSI for this task. The first method was provided by Stanford Research International (SRI) and is based on a two-class HMM decoder with a minimum duration of 30 ms (three frames) enforced with a three-state HMM structure trained on telephone conversations and further tuned to RT02s data. The features used in SRI's SNS detector, MFCC12, are different from the features used in the subsequent process of segmentation and clustering. The resulting speech segments are merged to bridge short nonspeech regions and padded according to NIST scoring guidelines. The SNS detector used was the same as that used in the RT05s evaluation. The parameters of the detector were tuned on the RT05s meeting development data to minimize the combination of misses and FAs reported by the NIST mdeval scoring tool [7] (more information about this can be found in Section 3).

The second SAD technique was developed at ICSI and does not require any outside data to train the system, although its parameters have been tuned with the Devel06 data set. The method is based on an iterative two-class segmenter that is initialized by considering all frames that fall below a certain relative energy threshold to be nonspeech. More information on this system can be found in [30]. Although the method has a similar SAD performance, the total DER is improved by using it. The method assumes that most nonspeech segments are silence with low background noise or close to it.

2.3 Iterative Segmentation and Clustering

The segmentation and agglomerative clustering process used was originally proposed by Ajmera and Wooters [18] and is shown in Fig. 2. The first module is the initialization, which will be explained later. Then, a resegmentation is carried out using Viterbi decoding with the initial HMMs along with a step to retrain the GMMs. This process may be iterated several times. Next, a cluster comparison and merging is carried out. When a merging takes place, the GMM for the new cluster is retrained with the data now assigned to it and the number of parameters (mixtures) of the merged model is the sum of the number of mixtures of the component models. The initial number of mixtures used for every cluster is a parameter that has to be determined empirically. A large number of initial mixtures may result (after cluster merging) in a high number of parameters that cannot be suitably trained. On the other hand, a small number of initial mixtures may result in poor modeling. The segmentation and clustering steps are repeated until a stopping criterion is reached. The segmenter consists of an ergodic HMM with a number of states equal to the number of speaker clusters (see Fig. 3). An individual cluster model consists of a set of substates, where the number of substates is determined by the minimum duration of each cluster. Every substate is modeled using a GMM containing a number of components that has to be specified initially. After passing through the minimum number of substates, the system can transition to a new cluster or stay in the same one. The transition is determined, in our system, solely by acoustics. This means that no penalty is applied at the final state of the cluster. If the log-likelihood of the last frame given a particular cluster is the largest of all the clusters, the systems stays in this cluster; otherwise, it transitions to the cluster that has the highest log likelihood. (see Figs. 3 and 4).

2.4 Merging and Stopping Criterion

One of the main problems in the segmentation and clustering process is deciding which merging and stopping criterion to use. The ΔBIC criterion has been used extensively, providing good results in Broadcast news data BN [17], [19], and the modification of ΔBIC to eliminate the penalty term has also given us good results for BN data [18]. The modified ΔBIC that we use for merging clusters is

$$\Delta BIC = \log p\Big(D\Big|\theta\Big) - \log p\Big(D_a\Big|\theta_a\Big) - \log p\Big(D_b\Big|\theta_b\Big), \quad (1)$$

[3] If there is more than one speaker in the reference, every missed speech error will be multiplied by the number of speakers.

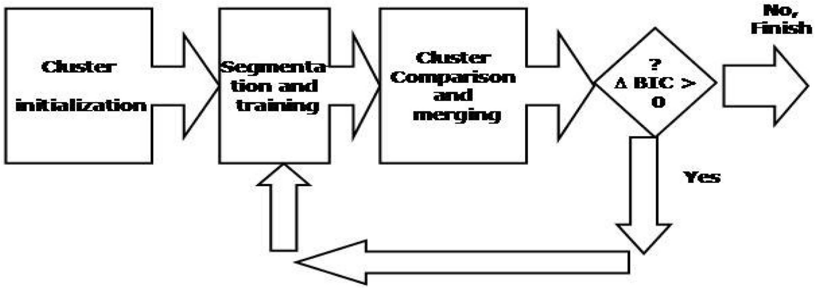

Fig. 2. Segmentation and clustering process.

where θ_a is the model created from D_a, θ_b is the model created from D_b, and θ is the model created from D, the union of D_a and D_b. The key to this modified BIC is that the number of parameters in θ must equal the sum of the number of parameters in θ_a and θ_b. Nevertheless, it is still an open question as to how much the performance depends on the kind of data vectors and models used in the comparisons, particularly if the number of parameters in the combined cluster has to be the sum of the number of parameters of the separate clusters (see [31]). If $\Delta\mathrm{BIC}$ is greater than 0 for a particular pair of clusters, those two clusters are believed to be similar enough to be merged. We find and merge the cluster pair that gives the largest $\Delta\mathrm{BIC}$. This basic method has been used at ICSI since 2003 to carry out speaker diarization with a single channel source such as the SDM condition.

2.5 Initialization

Ajmera and Wooters claim in their paper that the initialization procedure is not important to the diarization process [18]. They divided the data into K parts with equally long segments and these segments were used to train the initial GMMs. An additional loop of segmentation and training could be made before proceeding to the clustering module. Although other published results indicate that the initialization may be crucial to this process [22], [32], the objective in this paper was to study the influence of TDOA features in diarization not the influence of initialization. For this purpose, uniform segmentation has been used. The parameter K (the number of initial clusters) has to be decided empirically. If K is very small, there will be a high probability of missing some speakers since our method is bottom up (agglomerative). On the other hand, if K is very large, the system may stop at a very large number of clusters (speakers), increasing the error rate.

3 Data Used and Evaluation Metric

3.1 Data Sets

In this paper, we will use data coming from all of the NIST releases related to this task in the years 2002-2006: RT02s, RT04s, RT05s, and RT06s [29]. When the experiments are carried out with a subset of the data, it will be specified. A selection of data coming from RT02s, RT04s, and RT05s has been used as a special development set. This selection was made in our laboratory in order to fine-tune the algorithms that were going to be presented to the RT06s official evaluation campaign. We will call this selection the Devel06s set. This set is presented in Table 1. In general, the meetings consist of up to 16 speakers and up to 16 microphones located on top of a table or at a distance from the speakers. These microphones are considered "distant," which means that the microphone is not close to a speaker's mouth. In Table 1, we present all of the meetings used along with the number of microphones and speakers per meeting.

3.2 Evaluation Metric

The speaker diarization performance is evaluated by comparing the hypothesis segmentation, given by the system, with the reference segmentation provided by NIST [7]. This reference segmentation was generated by hand according to a set of rules also defined by NIST. In the

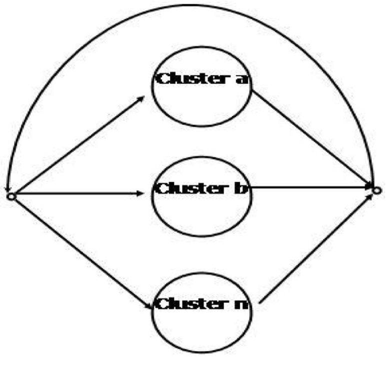

Fig. 3. Model for an entire meeting. A cluster tries to model a speaker and has the topology presented in Fig. 4. At the end of every cluster, a transition to any of the other clusters is permitted. This way, any speaker can take his turn after any other speaker. The nonspeech regions are previously removed (after Ajmera and Wooters).

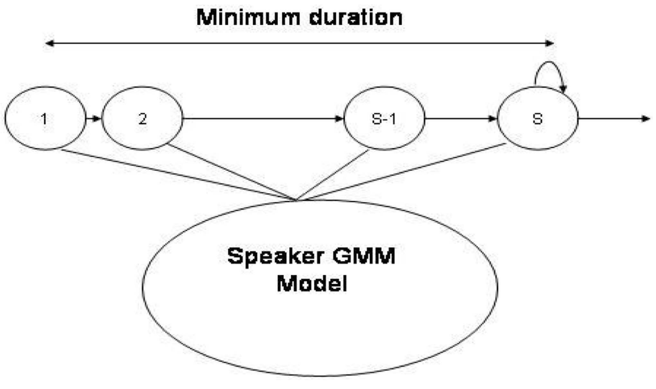

Fig. 4. Model for a cluster (from Ajmera and Wooters).

TABLE 1
Set of Meetings in the Databases Used with the Number of Microphones and Speakers per Meeting

RT05s	mic	speakers	DEVEL06S	mic	speakers	RT06s	mic	speakers
AMI_20041210-1052	12	4	AMI_20041210-1052	12	4	CMU_20050912-0900	2	4
AMI_20050204-1206	16	4	AMI_20050204-1206	16	4	CMU_20050914-0900	2	4
CMU_20050228-1615	3	4	CMU_20050228-1615	3	4	EDI_20050216-1051	16	4
CMU_20050301-1415	3	4	CMU_20050301-1415	3	4	EDI_20050218-0900	16	4
ICSI_20010531-1030	6	7	ICSI_20000807-1000	6	6	NIST_20051024-0930	7	9
ICSI_20011113-1100	6	9	ICSI_20010208-1430	6	7	NIST_20051102-1323	7	8
NIST_20050412-1303	7	7	LDC_20011116-1400	8	3	VT_20050623-1400	4	5
NIST_20050427-0939	7	4	LDC_20011116-1500	8	3	VT_20050623-1400	4	4
VT_20050304-1300	2	5	NIST_20030623-1409	7	6			
VT_20050318-1430	2	5	NIST_20030925-1517	7	4			
			VT_20050304-1300	2	5			
			VT_20051027-1430	2	5			

evaluation plan, the evaluation metric and a program to calculate it from both transcriptions is also defined. The error obtained is called the DER and it takes three errors into account (miss, FA, and speaker error). The error is time based. A miss error occurs when a speech segment is classified as nonspeech or an overlapping speaker is missing in the hypothesis. An FA error occurs when the system produces a speaker hypothesis when there is no speech in the reference. To calculate the speaker error, the program maps the hypothesis speakers to the reference speakers (only one reference speaker to one hypothesis speaker) in an optimal way so the overlap in duration between all pairs of reference and hypothesis speakers is maximized. A speaker error occurs for any region in the hypothesis that is mapped to a wrong speaker in the reference.

Because the metric is time based, it is weighted toward the loquacious speakers. An error for a speaker who does not speak much is less important than an error for a loquacious speaker. Consequently, the DER obtained for a specific meeting may be very much dependent on how many speakers talk and the time relationship between loquacious and nonloquacious speakers.

For the purpose of this paper, two kinds of reference transcriptions have been used. The first ones are the official hand-made references delivered by NIST. The second ones were created by aligning the official textual transcriptions obtained from the individual headset microphones with the ICSI-SRI speech to text system presented to the RT05s evaluation [24]. A reason for using force-aligned labels was that the RT06 evaluation campaign was originally intended to use those labels, but, later, this condition was dropped and hand-made labels were used. One question that arose in the last evaluation campaign was the appropriateness of using hand labels to compare and evaluate systems, especially when overlapping speech is included in the evaluation. During this year's development period, we experienced difficulties when using hand-made reference files, mostly when scoring on speaker overlapping regions. By comparing the hand-made references with the acoustic data, we observed that varying amounts of extra padding were inserted around each speaker overlap region, making its duration much longer than the actual acoustic event. We also observed some speaker overlapping labels on non-speaker-overlapping regions—because the hand references were created with close talk microphones, the overlap may be noticed by the labelers who were listening to the Individual Head Mounted (IHM) microphone channels, but the overlap is masked by noise in the MDM channels. All of these artifacts create an extra amount of missed-speech and speaker error that is not consistent over the different evaluation sets. Therefore, for the 2006 system development we decided to use references derived from forced alignments. Results with force-aligned labels were also calculated in the RT06 evaluation campaign by NIST, although they were not considered official [28].

4 Speaker Diarization Using Acoustic Features

The signals coming from the different microphones are Wiener filtered to improve the SNR as in previous systems [33]. Then, one of the signals (microphones), the one with the highest SNR, is selected as a reference channel. The TDOA between each of the other channels and the reference channel is calculated.

4.1 Time-Delay Calculation

In order to estimate the TDOA between segments corresponding to two microphones, we used a modified version of the Generalized Cross Correlation with phase transform ($GCC_{PHAT}(f)$) [34]. $GCC_{PHAT}(f)$ has been used by several people in the blind signal separation field [35]. The N best peaks are calculated and the best is selected with a postprocessing mechanism, see [26].

4.2 Acoustic Fusion

Once the delays are calculated every 500 ms (with a window shift of 250 ms), the signals are delayed and added together with a triangular window to generate a new composed signal (beamformed signal). The composed signal is then processed as a single signal. MFCC of 19th order are calculated every 30 ms using a window shift of 10 ms. These vectors are used in the segmentation and agglomerative clustering process (only the ones corresponding to the speech part). Using this procedure on the RT05s set, we obtained an 18.48 percent DER (using the

TABLE 2
Speaker Diarization Errors DER for the RT05s MDM Conference Room Evaluation Set

	# of initial clusters K	
Initial # of mixtures per cluster	**10**	**20**
1	31.20 %	34.77 %
2	38.68%	43.49%

standard NIST scoring software without counting overlaps). It is worth mentioning that, for this experiment, we used an initial number of clusters of 10, an initial number of Gaussian mixtures per model of 5, and a minimum duration of a cluster of 3s.

5 Speaker Diarization Using Only Between Channel Differences

5.1 Baseline

For the speech regions, we calculate the TDOAs using the procedure mentioned in the previous section. The window shift is 10 ms, the same as the one used for the calculation of acoustic features. We form a vector of delays that has as many components as the number of microphones minus 1. Nonspeech frames, estimated previously, are excluded from the subsequent process. The vector of delays is then fed into the aforementioned segmentation and agglomerative clustering module instead of the acoustic vectors [36]. We experimented with several values for the segmentation and agglomerative clustering parameters such as the initial number of mixtures per cluster and the number of initial clusters as mentioned in the previous section. In contrast to the insensitivity to the parameters when using acoustic vectors as mentioned by Ajmera and Wooters [18], there is sensitivity to the parameters when we used only delay vectors. In Table 2, we present the DER for different sets of parameters for the RT05s data set using a minimum duration of 2 sec.

In the subsequent experiments, we will be using one initial mixture and 10 initial clusters. Obviously, if the number of speakers in the room is more than 10, the errors of the system will dramatically increase. In Table 3, we present the DER for RT05s and the components of it (Miss error, FA error, and Speaker error). Note that the SNS error is the addition of the Miss error plus the FA error. It is not surprising that one initial mixture gives better results than two initial mixtures because, if we assume that the speakers do not move far from their seats, the information contained in the delay vector is likely to be unimodal. In any event, due to errors in the delay calculations and some small movement of speakers, the real distribution of the data may be multimodal. This fact is automatically modeled in our system because, when two clusters are merged, the number of mixtures of the merged model is the sum of the number of mixtures of the component cluster models.

In Fig. 5, we present the DER for every meeting, comparing the results using only the acoustics and the results using only the TDOAs. The results using only the delays are less stable across different meetings than the results using the acoustics and the average results using acoustics only are better than the ones using delays only. It can also be seen that there is a set of meetings whose results for every method are very similar. One can also see that the results for a pair of meetings are extremely bad. This effect may be due to several factors, such as the total number of speakers and total number of turns. In [37], there is a study on the diarization results for several shows in the BN domain across several algorithms. The authors conclude that there are shows that are very difficult to analyze (they called them "nuts") and others have a large amount of variation in the DER when using different algorithms (they called them "flakes"). This fact is also demonstrated in our data, as can be seen in Fig. 5. Some meetings perform poorly and others show a large variation in DER across different algorithms (acoustic vectors or delay vectors). The result of their study is that the shows that are more difficult to diarize are the ones that contain many speakers, many speaker turns, and the absence of a dominant speaker.

In order to compare our results with those presented by Elis and Liu [27], we have run the system with the same set of meetings that they used in their experiments and have reduced the number of channels available to four in all cases (Elis and Liu used only four channels). The comparisons of both experiments are presented in Table 4. It is important to note that, in these results, two of the meetings from the NIST RT04s development data (the CMU meetings) have not been used because they contain only one distant microphone

TABLE 3
Missed Speech, FA Speech, Speech/Nonspeech Error Speaker Error and Diarization Error for the RT05s Evaluation Set Using One Mixture and 10 Initial Clusters

File	Miss	FA	S/NS	Spkr	Total
AMI_20041210-1052	1.1	1.9	3	13,5	16.53
AMI_20050204-1206	1.8	1.7	3.5	19.6	23.03
CMU_20050228-1615	0.1	1	1.1	17.2	18.28
CMU_20050301-1415	0.2	3.3	3.5	42.4	45.88
ICSI_20010531-1030	4.3	1.3	5.6	15	20.59
ICSI_20011113-1100	2.9	2.7	5.6	39.9	45.52
NIST_20050412-1303	0.6	2.9	3.5	21.7	25.19
NIST_20050427-0939	1.5	2.5	4	33.2	37.18
VT_20050304-1300	0	3.6	3.6	22.1	25.7
VT_20050318-1430	0.3	22.6	22.9	38.4	61.27
ALL	1.3	4	5.3	25.9	31.2

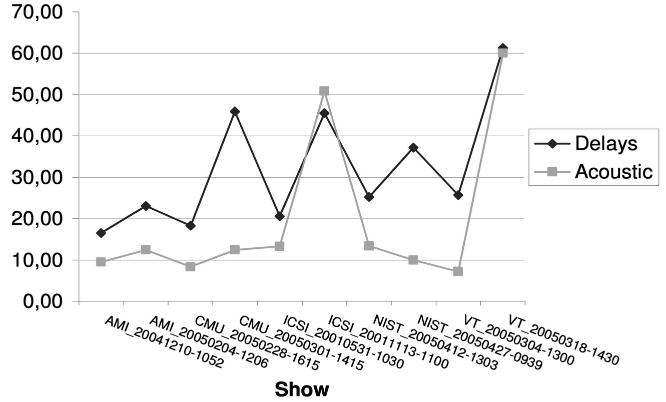

Fig. 5. Results across RT05s meetings for two different systems: using acoustics only or delays only.

and are thus not compatible with the conditions of our experiment (MDM).[4] The results presented here also include the overlapping regions and no FAs (we call it the DER* error). We have also included the standard DER error in Table 4 (including the FAs) for completeness. The analysis of the results shows a large improvement compared to that of Elis and Liu. The differences may well come from the different ways of calculating the delays between signals and the different segmentation and clustering procedure. Since the number of microphones used in this experiment was less than the number of microphones available, we have also computed the error rate that we could obtain for the same set of meetings if we had used all of the available microphones. Table 5 shows the results of this comparison. It can be seen that the use of more microphones reduces the DER error rate by 8.8 percent relative.

5.2 Delay Calculation Improvements

The calculation of the delays is not exact and, occasionally, the autocorrelation between signals does not find the highest peak for the correct delay value. In this paper, we have also experimented with a new method to make the calculation of the delays more robust [38]. It consists of two phases. The first phase processes N peaks (experiments carried out with $\mathrm{N} = 8$) of the cross-correlation between each channel and the reference and performs a Viterbi alignment to extract the two best paths (two best peaks) for every frame. For the Viterbi process, the emission probability used is the cross-correlation value for the peak and the transition probability between two nodes (frames) is the inverse of the difference between delay values, ensuring that the N transition probabilities in a particular instant sum to 1. The second phase processes the two best peaks for every frame and every channel and performs a Viterbi alignment between all channels to find the best path of the delay vector across the entire sentence. The emission probabilities are the product of the individual correlation values of each delay. The transition probabilities are computed by adding all delay distances from all considered delays, normalized to sum to 1. This technique aims to find the optimum trade-off between reliability (value of the cross correlation) and stability (distance between delays corresponding to contiguous frames).

TABLE 4
Comparisons between Results Obtained by Elis and Liu and Our Results in the Same Subset of Meetings from NIST RT04 Development Data-DER^{ast}

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

The DER for the Devel06s set using delays obtained by the baseline system is 35.39 percent and the DER obtained using the improved system is 29.45 percent. Thus, for this set of meetings, we can see that the improvement that we obtained by using the improved method is 16.78 percent relative. In contrast to the results presented so far, the DER presented in this section was evaluated using overlapping regions, force-aligned transcriptions, and the second ICSI-developed SAD technique.

6 Speaker Diarization Mixing Between Channel Differences and Acoustic Parameters

After having experimented with acoustic vectors only and delay vectors only, the obvious continuation is to combine them [39]. The first idea that we had was to concatenate both vectors, that is, join the MFCC vectors and the TDOA vectors in a single vector, but we could not obtain an improvement compared to the use of acoustic vectors only. We believe that this was due to the use of diagonal covariance matrices to model the multidimensional Gaussians. Another possible reason for this method not working is the fact that the number of initial Gaussians used in the MFCC models alone is five compared to one in the case of TDOA models (parameters determined empirically, see the Section 5). The inherent nature of the data (TDOAs for a speaker in a determined location tend to be unimodal,

TABLE 5
Comparisons between DER Obtained Using Four Channels and Results Using All of the Channels Available in the System

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

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

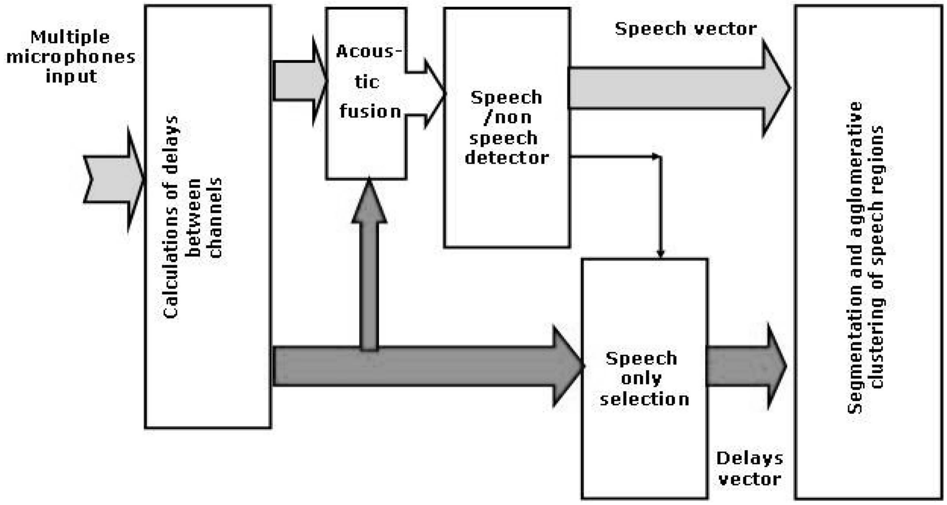
Multiple microphones input
Calculations of delays between channels
Acous-tic fusion
Speech /non speech detector
Speech vector
Speech only selection
Delays vector
Segmentation and agglomerative clustering of speech regions

Fig. 6. General architecture of the system that uses both acoustic and delay vectors.

whereas MFCCs for a speaker tend to be multimodal) led us to develop a different method of combining both vectors. Therefore, we decided to keep both vectors separate and model the clusters with independent information coming from both sets of vectors. The general architecture of the system is detailed in Fig. 6.

The channels are first processed to obtain the delays between them and create both a beamformed signal and a vector of delays. The beamformed signal is used to classify speech versus nonspeech. The speech regions are then processed to obtain the MFCC, as mentioned in Section 4. This set of vectors is used in parallel with the delay vectors by the segmentation and agglomerative clustering module. The segmentation process uses the log likelihood of the best path to create a segmentation hypothesis. The agglomerative clustering uses ΔBIC to define the clusters to merge, which also requires the computation of the log likelihood of a set of vectors given a model. For the combined system, we used a joint log likelihood as follows:

$$\log p(x[n], y[n]|\theta_a) = \alpha \log p(x[n]|\theta_{ax}) + (1-\alpha)\log p(y[n]|\theta_{ay}). \qquad (2)$$

θ_{a} is the compound model for any given cluster a, θ_{ax} is the model created for cluster a using the acoustic vectors x[n], and θ_{ay} is the model created for cluster a using the delay vectors y[n]. α is a weight factor that has to be determined. The DER for the RT05s set (not counting overlaps) as a function of the weighting factor used is presented in Fig. 7. Starting from a DER of 31.2 percent for delays only and 18.48 percent for acoustic only, we obtain 15.46 percent for the compound system using a weight of 0.9 (using a minimum duration of 2 sec and 10 initial clusters). This is a DER reduction of 16.34 percent relative. In Fig. 8, we show the same plot, this time for the Devel06s set. It is important to mention that, in this plot, compared to the previous one, overlap and force-aligned labels were used in scoring. Also, for this experiment, the speech/nonspeech detector was changed and the system described in [30] was used. Finally, for this experiment, we used the improved method for delay calculations, a minimum duration of 2.5 sec and 16 initial clusters. From 29.45 percent using only delays and 13.44 percent using only acoustic features, we obtain a DER of 9.74 percent, also using a weight factor of 0.9. This means a DER reduction of 27.52 percent. In Table 6, we give details of the DER obtained for each meeting, for future reference.

In Table 7, in the first and third columns, we present the aforementioned data. In the second column, we present the results obtained using the old delay calculations (called System A). The improvement in the delay calculations for the combined system is 3.4 percent relative (see the third column

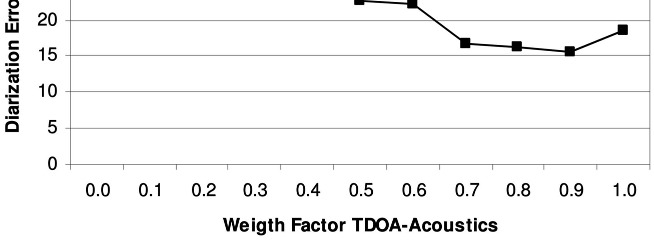

Fig. 7. Plot of DER as a function of the weight factor applied for the RT05s.

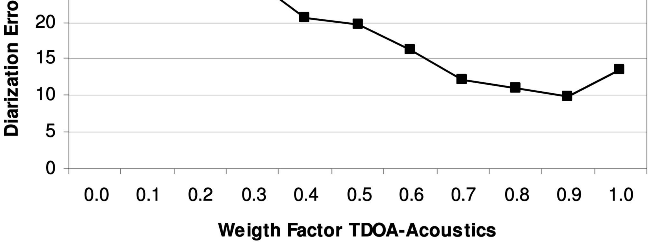

Fig. 8. DER as a function of the weight factor for the Devel06s data.

TABLE 6
Results for the Set Devel06s

	Miss	FA	Spkr	Total
AMI_20041210-1052	0.40	1.40	1.80	3.48
AMI_20050204-1206	2.40	3.10	2.30	7.75
CMU_20050228-1615	9.10	1.00	2.40	12.51
CMU_20050301-1415	3.40	1.60	2.20	7.13
ICSI_20000807-1000	4.60	0.40	2.90	7.94
ICSI_20010208-1430	3.30	1.00	11.10	15.42
LDC_20011116-1400	2.10	2.90	0.50	5.53
LDC_20011116-1500	5.90	1.10	6.50	13.58
NIST_20030623-1409	1.00	0.70	0.80	2.48
NIST_20030925-1517	7.10	5.50	3.60	16.17
VT_20050304-1300	0.60	1.00	5.30	6.91
VT_20050318-1430	1.40	7.40	17.50	26.32
All	3.30	2.10	4.40	9.74

We present the percentage of missed speech, FA speech, speaker error, and total diarization error (DER).

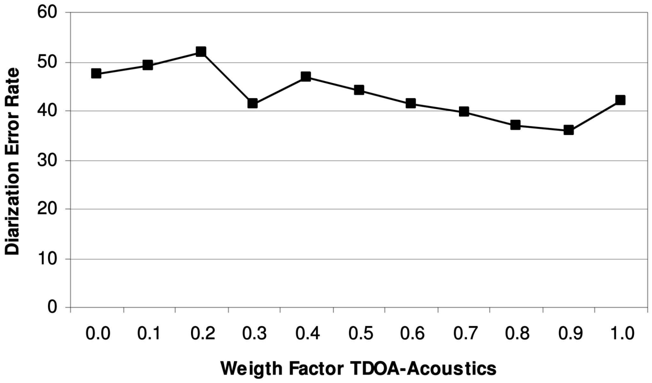

Fig. 9. DER as a function of the weight factor for the RT06s data.

in Table 7, System B). In the fourth column, we present the results obtained from this system (weight factor 0.9) at the official RT06s evaluation campaign (35.77 percent DER, which was a top performer [28]). In the same column, we also present the results obtained after the evaluation with acoustic only data and delay only data, resulting in a 15.1 percent relative improvement over the acoustic only result.

In Table 7, we also present the data obtained with force-aligned labels (fifth column). The relative improvement (acoustic only versus acoustic plus delays) obtained using force-aligned labels is larger (25.84 percent).

If we examine Table 7, we can see that the relative improvement in the combined system compared to the acoustic only system is greater when we use force-aligned labels even though the base error is lower—25.84 percent versus 15.10 percent. We can also see that the relative improvement in the Devel06s data set, 27.52 percent (force-aligned data), is greater than the improvement obtained in the RT05s 16.34 percent (hand aligned). We believe that force-aligned labels provide a better reference to compare different algorithms than hand-aligned labels.

In Fig. 9, a plot of the results obtained after the evaluation using the RT06s data is shown. It can be seen that, fortunately, the minimum DER is obtained at the same point (weight factor 0.9) as in our development data. However, for weight factors 0.1 and 0.2, the combined error is greater than with the TDOA or the MFCC values alone. We did not find this behavior in our development data, as seen in Figs. 7 and 8. This maximum comes exclusively from the EDI_20050216-1051 and EDI_20050218-0900 shows. The EDI site is a new site that is not included in either RT05s or Devel06s. One possibility is that the joint information of acoustic and delays counteract each other, giving worse results. It is also possible that the existence of overlapping regions or the movement of the speakers has an influence, but this should be investigated further (although we could analyze the amount of overlapping speakers in RT06s, we do not have data on the location across time of the speakers or that of the microphones). In Fig. 10, a breakdown of DER for every meeting in RTO6s is presented.

7 Discussion

In Section 5, we mentioned the problem that some shows give good DER results and others give much worse results. The method that we have used to segment the speakers using delays only assumes that the speakers are not moving far from their positions. In other words, the system assigns a speaker to a region in space. If this assumption is invalid, the DER using TDOAs only will be severely increased. Alternatively, using only acoustics, some shows give poor performance and one of the reasons may be the existence of

TABLE 7
DER for the Eval05s and Devel06s and the Official RT06s Data Set Obtained
Using Acoustic Features Only, Delay Features Only, and Combined Features

Features used	RT05s (hand labels-no overlap-old SNS) System A	Devel06s (force-aligned labels-overlap-new SNS) System A	Devel06s (force-aligned labels-overlap-new SNS) System B	RT06s (hand labels-overlap-new SNS) System B	RT06s (force-aligned labels-overlap-new SNS) System B
Delays only	31.20 %	35.39 %	29.45%	47.55% (after the eval)	36.99% (after the eval)
Acoustic features only	18.48 %	13.44 %	13.44%	42.13% (after the eval)	27.01% (after the eval)
Combined acoustic+delays	15.46 %	10.09 %	9.74%	35.77% (official result)	20.03% (after the eval)
Relative error reduction	16.34 %	21 %	27.52%	15.10% (after the eval)	25.84% (after the eval)

System A stands for original delay calculations and System B for improved delay calculations.

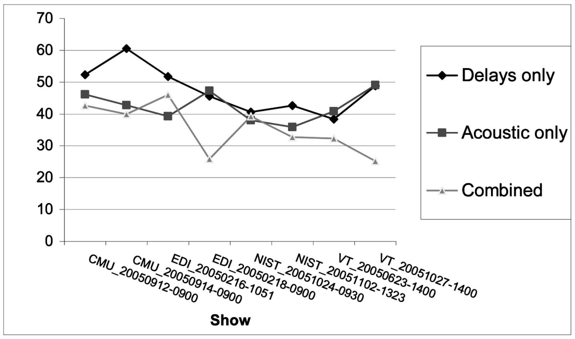

Fig. 10. Breakdown of DER for every meeting in RT06s. Results are presented using acoustics only, delays only, and the combined system.

speakers who are close in the acoustic space. An advantage to mixing delay information with acoustic information is that it results in a system that is robust against the weaknesses in either one or the other dimension.

In the RT06s evaluation, we used a weight factor between delays and acoustics, which was optimum according to our development set. However, again, the optimal weight factor may be dependent on the meeting itself. To illustrate this problem, in Fig. 11, we present the DER across different weights on the Devel06 set (labeled Average all) versus the same plot for a subset of the Devel06 set (labeled Average subset) containing just the CMU_20050228-1615, LDC_20011116-1500, and VT_20050304-1300 meetings. For this experiment, the only change that we made was the weight factor. We can see that the minimum DER for the subset of meetings appears when the weight factor is 0.7. That may well correspond to the case of several speakers who are acoustically close but who are in fixed well-separated areas of the room. In Fig. 12, we present a breakdown per set of meetings for the RT06s set. We notice that there is a set of meetings with minimum weight of 0.7, another set with a minimum weight of 0.8, and there is even a meeting with a minimum weight of 0.3. The overall minimum weight is 0.9, as shown in Fig. 9.

The results presented here could be further improved by improving the discrimination capability of both methods separately. We have shown that, by using the improved method of calculating the delays presented in Section 6, we have been able to get better results for the Devel06s set and decrease the DER by 3.4 percent relative. These results could be further improved if we add other sources of information, such as pitch, as preliminary experiments carried out in our laboratory have demonstrated [40].

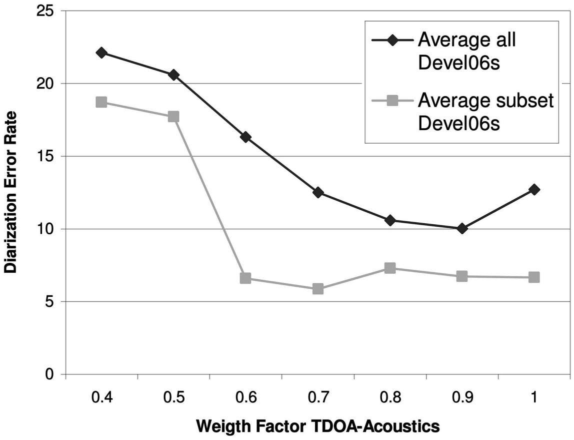

Fig. 11. Comparison of DER across different weight factors for the Devel06s set (average all) and for a subset of it (average subset).

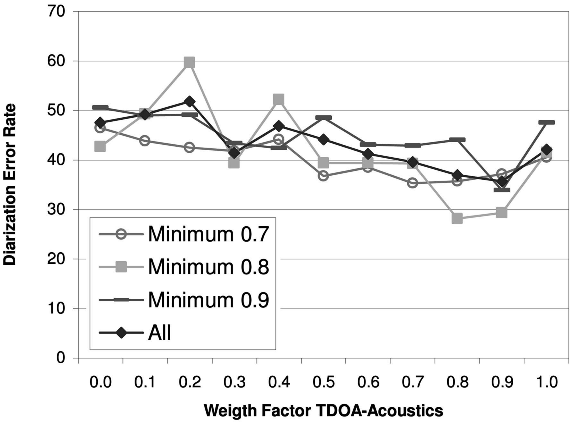

Fig. 12. Analysis of DER for RT06s data depending on the set of meetings. The label Minimum 0.7 corresponds to the average from the CMU_20050914-0900, NIST_20051024-0930, and VT_20050623-1400 meetings, whose minimum weight is 0.7, the label Minimum 0.8 corresponds to the average from the EDI_20050218-0900 and NIST_20051102-1323 meetings, the label Minimum 0.9 corresponds to the average from the CMU_20050912-0900 and VT_20051027-1400 meetings, and the label All corresponds to the average from all of the meetings. The EDI_20050216-1051 meeting has a minimum at 0.3 and is one of the meetings with a maximum at $\mathrm{weight} = 0.2$.

8 Conclusion

We have proposed and developed a new method to mix delay information from different channels with acoustic information to improve the task of speaker diarization for meetings with multiple distant microphones. The results are encouraging and a first step on the path of combining as many sources of information as possible to solve the problem. Of particular interest could be the inclusion of suprasegmental information, such as pitch, language models, and so forth, and other techniques used in speaker verification/recognition systems. An important area of research is the development of a robust mechanism to combine all sources of information that is stable against diverse shows and application environments. Another relevant area of research would be to include some of the techniques developed here in an online system to make discriminative and interactive communication between humans and computers in meetings an attainable goal.

Acknowledgments

This work was supported by the Joint Spain-ICSI Visitor Program and by the projects ROBINT (DPI 2004-07908-C02), TINA (UPM-CAM R05-10922), and EDECAN (TIN2005-08660-C04). The authors would like to thank Andreas Stolcke, Kemal Sönmez, and Nikki Mirghafori for many helpful discussions and James Fung for his revision of the paper. They also thank the anonymous reviewers for their comments that helped to improve the final paper. This work was carried out while José M. Pardo was a visiting fellow at

the International Computer Science Institute, Berkeley, California.

REFERENCES

[1] http://nist.gov/speech/tests/rt/rt2002/, 2007.

[2] J. Ajmera, G. Lathoud, and I.A. Mc Cowan, "Clustering and Segmenting Speakers and Their Locations in Meetings," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '04)*, pp. I-605-I-608, 2004.

[3] G. Lathoud, I.A. Mc Cowan, and J.M. Odobez, "Unsupervised Location-Based Segmentation of Multi-Party Speech," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '04) —NIST Meeting Recognition Workshop*, May 2004.

[4] G. Lathoud and I.A. Mc Cowan, "Location Based Speaker Segmentation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '03)*, pp. I-176-I-179, 2003.

[5] G. Lathoud, "Further Applications of Sector-Based Detection and Short-Term Clustering," IDIAP RR 06-26, May 2006.

[6] D.A. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. V-953-V-956, 2005.

[7] NIST Spring 2006 (RT06s) Rich Transcription Meeting Recognition, http://www.nist.gov/speech/tests/rt/rt2006/spring/, 2007.

[8] S.E. Tranter and D.A. Reynolds, "Speaker Diarisation for Broadcast News," *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pp. 337-344, May 2004.

[9] C. Barras, X. Zhu, S. Meignier, and J.L. Gauvain, "Improving Speaker Diarization," *Proc. DARPA 2004 Rich Transcription Workshop (RT '04)*, Nov. 2004.

[10] S. Meignier, D. Moraru, C. Fredouillea, J.-F. Bonastre, and L. Besacier, "Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization," *Computer Speech and Language*, vol. 20, pp. 303-330, Apr.-July 2006.

[11] S. Tranter and D.A. Reynolds, "An Overview of Automatic Speaker Diarization," *IEEE Trans. Audio, Speech and Language Eng.*, vol. 14, no. 5, pp. 1557-1565, Sept. 2006.

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

[13] X. Zhu, C. Barras, S. Meignier, and J.L. Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization," *Proc. European Conf. Speech Comm. and Technology*, pp. 2441-2444, Sept. 2005.

[14] D.A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations," *Proc. Fall Rich Transcription Workshop (RT '04)*, 2004.

[15] R. Sinha, S.E. Tranter, M.J.F. Gales, and P.C. Woodland, "The Cambridge University March 2005 Speaker Diarization System," *Proc. European Conf. Speech Comm. and Technology*, pp. 2437-2440, Sept. 2005.

[16] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and L. Bonastre, "The ELISA Consortium Approaches in Broadcast News Speaker Segmentation during the NIST 2003 Rich Transcription Evaluation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '04)*, 2004.

[17] J. Ferreiros and D. Ellis, "Using Acoustic Condition Clustering to Improve Acoustic Change Detection on Broadcast News," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '00)*, 2000.

[18] J. Ajmera and C. Wooters, "A Robust Speaker Clustering Algorithm," *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU '03)*, 2003.

[19] S.S. Chen and P.S. Gopalakrishnan, "Speaker Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Feb. 1998.

[20] http://www.nist.gov/speech/tests/rt/rt2003/spring/, 2007.

[21] http://www.nist.gov/speech/tests/rt/rt2004/fall/, 2007.

[22] C. Wooters, N. Mirghafori, A. Stolcke, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf, "The 2004 ICSI-SRI-UW Meeting Recognition System," *Proc. Joint AMI/Pascal/IM2/M4 Workshop Meeting Recognition*, 2005.

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

[24] 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," *Proc. NIST MLMI Meeting Recognition Workshop*, 2005.

[25] 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," *Proc. NIST Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '04)—Meeting Recognition Workshop*, 2004.

[26] X. Anguera, C. Wooters, and J. Hernando, "Speaker Diarization for Multi-Party Meetings Using Acoustic Fusion," *Proc. IEEE Workshop Automatic Speech Recognition and Understanding (ASRU '05)*, 2005.

[27] D.P.W. Elis and J.C. Liu, "Speaker Turn Segmentation Based on Between-Channels Differences," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '04)*, 2004.

[28] J.G. Fiscus, J. Ajot, M. Michel, and J.S. Garafolo, "The Rich Transcription 2006 Spring Meeting Recognition Evaluation," *Lecture Notes in Computer Science*, vol. 4299, p. 319, http://www.nist.gov/speech/tests/rt/rt2006/spring/pdfs/rt06s-SPKR-SAD-results-v5.pdf, 2006.

[29] http://www.nist.gov/speech/tests/index.htm, 2007.

[30] X. Anguera, M. Aguiló, C. Wooters, C. Nadeu, and J. Hernando, "Hybrid Speech/Non-Speech Detector Applied to Speaker Diarization of Meetings," *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, June 2006.

[31] X. Anguera, C. Wooters, and J. Hernando, "Automatic Cluster Complexity and Quantity Selection," *Lecture Notes in Computer Science*, vol. 4299, pp. 248-256, 2006.

[32] X. Anguera, C. Wooters, and J. Hernando, "Friends and Enemies: A Novel Initialization for Speaker Diarization," *Proc. Int'l Conf. Spoken Language Processing*, 2006.

[33] N. Mirghafori et al., "From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '04)*, Oct. 2004.

[34] M.S. Brandstein and H.F. Silverman, "A Robust Method for Speech Signal Time-Delay Estimation in Reverberant Rooms," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '97)*, 1997.

[35] M. Swartling et al., "Direction of Arrival Estimation for Multiple Speakers Using Time-Frequency Orthogonal Signal Separation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. IV, pp. 833-836, 2006.

[36] J.M. Pardo, X. Anguera, and C. Wooters, "Speaker Diarization for Multi-Microphone Meetings Using Only Between-Channel Differences," *Lecture Notes in Computer Science*, vol. 4299, pp. 257-264, May 2006.

[37] N. Mirghafori and C. Wooters, "Nuts and Flakes: A Study of Data Characteristics in Speaker Diarization," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '06)*, 2006.

[38] X. Anguera, C. Wooters, and J.M. Pardo, "Robust Speaker Diarization for Meetings: ICSI RT06s Meetings Evaluation System," *Lecture Notes in Computer Science*, vol. 4299, pp. 346-358, 2006.

[39] J.M. Pardo, X. Anguera, and C. Wooters, "Speaker Diarization for Multiple Distant Microphone Meetings: Mixing Acoustic Features and Inter-Channel Time Differences," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '06)*, pp. 2194-2197, 2006.

[40] A. Gallardo-Antolín, X. Anguera, and C. Wooters, "Multi-Stream Speaker Diarization Systems for the Meetings Domain," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '06)*, Sept. 2006.

José M. Pardo received the degree in telecommunication engineering and the PhD degree from the Universidad Politécnica de Madrid in 1978 and 1981, respectively. He won a National Award in 1980 for the best graduate in telecommunication engineering and a National Award for the Best PhD Thesis in 1982. Since 1978, he has worked in speech technology and has held different teaching and research positions at the Universidad Politécnica de Madrid. He has been the head of the Speech Technology Group since 1987 and a full professor since 1992. He was head of the Electronic Engineering Department from 1995 to 2004. He was a Fulbright Scholar at the Massachusetts Institute of Technology (MIT) in 1983-1984, a visiting scientist at SRI International in 1986, and, recently, a visiting fellow at the International Computer Science Institute in 2005-2006. He was a member of the International Speech Communication Association (ISCA) Advisory Council from 1996 to 2006. He was chairman of the European Conference on Speech Communication and Technology (EUROSPEECH '95) and member of the ELSNET Executive Board in 1998-2004. He has been a member of NATO RSG 10 and IST 3 since 1994. He is a senior member of the IEEE and a member of the Acoustical Society of America (ASA), ISCA, and EURASIP. He has authored or coauthored more than 160 papers and holds two patents.

Xavier Anguera received the MS degree and the PhD degree from UPC University in 2001 and 2006, respectively, with a thesis titled "Robust Speaker Diarization for Meetings." From September 2004 to September 2006, he visited the International Computer Science Institute (ICSI), where he worked on speaker diarization for meetings and participated in several NIST RT evaluations. He is currently with Telefónica I+D pursuing research and actively participating in Spanish and European projects. His research interests cover the areas of speaker recognition and automatic indexing of acoustic data. He is a member of the IEEE.

Charles Wooters received the BA and MA degrees in linguistics from the University of California (UC), Berkeley. He received the PhD degree from UC Berkeley in "speech recognition" in November 1993. This interdisciplinary program spanned the Departments of Computer Science, Linguistics, and Psychology. After graduating from UC Berkeley, he went to work for the US 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, he developed the speech recognition software systems that were used in Aesop and Hermes. In April 1997, he 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, he joined the speech group at the International Computer Science Institute in Berkeley, where he continues to perform research, specializing in automatic speaker diarization. He is a member of the IEEE.