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

José M. PARDO, *Senior Member IEEE,* Xavier ANGUERA, and Chuck WOOTERS

Abstract-- **Human-Machine interaction in meetings requires the localization and identification of the speaker that interacts with the system and the recognition of the message spoken. A seminal phase towards this goal is the so-called rich transcription research, which covers speaker diarization together with the annotation of sentence boundaries and elimination of speaker disfluencies. The subarea of speaker diarization intends 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 different microphones and propose an improved method to do speaker diarization for meetings with multiple distant microphones. The proposed algorithm makes use of acoustic information and information about the delays between signals coming from all the sources. With this procedure, ~~we have been able to achieve the best performance in the last spring 2006 National Institute of Standards Rich Transcription Evaluation,~~ improving from 15% to 20% relative the Diarization Error Rate (DER) compared to previous systems.**

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

I. INTRODUCTION

Human-Machine interaction in meetings requires the localization and identification of the speaker that interacts with the system and the recognition of the message spoken. A seminal

phase towards this goal is the so-called Rich Transcription research, which covers speaker

diarization together with the annotation of sentence boundaries and elimination of speaker

disfluencies. The Rich Transcription Research Area has been initially motivated when trying to

solve the problem of speech transcription for audio sources increasingly complex, telephone

conversations, broadcast news data or meetings data. The solution requires that the information is

annotated with as much detail as possible concerning speaker turns, sentence units etc. if a

posterior efficient use of the data is required (e.g indexation, translation etc.). With this ambitious objective in mind, the US National Institute of Standards and Tecnhology (NIST) started several years ago a series of evaluations taking this problem into account and defining the concept of Rich Transcription ~~as an alternative~~ to Speech to Text translation or Speech Recognition [1]. One of the tasks defined by NIST was speaker diarization. Speaker diarization for meetings is the task of identifying the number of participants in a meeting and ~~create~~ a list of speech time intervals for each such participant. It is important to note that the task is done without using any knowledge about the location or identity of the speakers in the room , the location and quality of the microphones, or the details of the acoustics of the room.

What are the uses of speaker diarization? In the first place, speaker diarization serves the purpose of aiding the transcription task. Instead of just transcribing a recording into unorganized text, the transcription is annotated also with a different label for each different speaker. Later if we knew also a set of possible speakers that could appear in the recording we could use a speaker verification algorithm and assign an identified speaker to every label. A transcription annotated in this manner is much more readable and usable than a transcription without this information. It could be also used for automatic speaker indexation of audio documents.

A second possible application of speaker diarization is to apply adaptation techniques to speech recognition. Once we know the regions that correspond to the same speaker, we could 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 [2].

Finally, the methods used to do speaker diarization (particularly if they use delays between signals) will surely be seminal in the more difficult task of identifying on line who is speaking in a meeting and what is being said, particularly if only multiple microphones are available (no

video cameras) and several speakers talk at the same time (there is overlap between them).

Since 2002, NIST has included speaker diarization as one of the task evaluated in the context of Rich Transcription of Meetings [12], which evolved from speaker diarization for broadcast news and telephone conversations. In 2002 the evaluation was carried on using a Single Distant Microphone (referred to as SDM). After 2004, the primary condition in the evaluation was to use Multiple Distant Microphones (MDM).

A. Speaker diarization for Meetings using a single distant microphone (SDM)

In general, the approaches used in the literature for speaker diarization using a single distant microphone or a single recording signal have their basis in previous audio segmentation and diarization applied to Broadcast News data [2],[18][19],[27],[28]. A good overview on this topic has been recently published [28].The usual process starts first eliminating non speech frames from the recording. This task sometimes is difficult since non-speech may include music, laughter, breath, lip-smack, paper shuffling etc. There are several alternatives to do this job. The first is just to use to Maximum likelihood classification with two GMM models, one for speech and one for silence and others [5]. Others model explicitly noise and music [30][31]. Finally speech detection can be made also using a phone recognizer or a word recognizer [32].Then the process follows by finding acoustic changes in the signal and create homogeneous segments. This is made by analyzing adjacent windows of data and calculating a distance between both [3]. In the integrated versions as the one used at ICSI there is no need to do this phase explicitly[4]. The next phase, specially for Broadcast News (optional) is to classify those segments into male-female, narrow-band /high-band. The homogeneous segments are later hierarchically clustered in order to joint 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 recluster the segments in an iterative process.

The LIA and CLIPS systems [27],[29], are a good example for the comparison of both approaches, step by step or integrated. The CLIPS system is a sequential system based on a speaker change detection followed by a hierarchical clustering. It uses the Global Likelihood Ratio (GLR) for acoustic change detection and as clustering distance and the BIC distance as a stop criterion. It uses also Maximum a Posteriori Adaptation to train the clusters from a background model. The LIA system on the contrary is an integrated approach that uses an evolutive Hidden Markov Model to generate speaker clusters top-down and retraining the data and resegmenting the show every time a new speaker is added to the model. Used separately, the LIA system outperforms the CLIPS system 16.9 % error compared to 19.3% error for the RT03s database. Both laboratories have created a joint system using the CLIPS as a first module followed by the LIA model as the second module obtaining 12.9% error and it ~~was the winner at the RT03s campaign.~~

Tranter and Reynolds[18] present in their paper also two systems, the one by CUED and the one by MIT-LL. The CUED system uses a step-by-step approach, but segmentation is not done with an acoustic change detector but with a phone recognizer. Clustering is done using as distance the Arithmetic Harmonic Sphericity and they compare three different stopping criteria, being the Bayes Information Criterion (BIC) the winner. The MIT-LL method is also a step by step method that uses as acoustic change detector a system based on adjacent window comparisons using the BIC criterion and a clustering and stopping criterion using also BIC. They also proposed in their paper a "plug and play combination" of the components of both systems giving the best DER a combination of CUED for segmentation and MIT-LL for clustering.

One of the best systems recently published for Broadcast News is the one presented by LIMSI [19]. Their system is an integrated system and they use innovative methods to improve the performance.They include a speaker identification module to do 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, and it uses training data in the same domain. ~~This system was the winner at the RT04f evaluation campaign~~.

The Meetings domain differs from the Broadcast News in that the forums are highly diverse, the participants have particular relationships and vocabularies, the meetings are highly interactive and there is simultaneous speech from multiple speakers. Furthermore distant microphones lead to reverberation speech and noise background. Consequently the problem is much more difficult than the Broadcast News domain. In 2002 NIST conducted a test of speaker diarization in the meetings domain in the SDM condition. Although the tests after 2002 considered MDM as the primary condition, the methods applied to SDM or previously to Broadcast News may be considered a first basic step toward the development of algorithms for MDM.

There has been extensive research at ICSI in the last few years in the area of meetings recognition including speech recognition and speaker diarization [3],[4],[5],[6],[7],[8],[9]. The basic method used at ICSI for SDM can be considered an integrated approach and models the utterance as an an ergodic HMM that has a number of states equal to the initial number of clusters or speakers (K). Each state in the HMM contains a sequence of MD substates which are used to impose a minimum duration to the cluster. Within a state, each one of the sub-states shares a probability density function (PDF) modelled with a Gaussian mixture model (GMM) with a diagonal covariance ~~matrix~~ [4],[5]. Essentially, the process consists of two modules: the initialization and the integrated segmentation and clustering. The initialization requires a "guess"

at the maximum number of speakers (K) that are likely to occur in the data. The data are then divided into K equal-length segments, and each segment is assigned to one model. Each model's parameters are then trained using its assigned data. These are the models that seed the posterior agglomerative clustering and segmentation processes in an iterative loop.

B. Speaker diarization for Meetings with multiple distant microphones

The task of speaker diarization for Meetings with multiple distant microphones (MDM) should be easier compared to the use of a single distant microphone (SDM) because: a) there are redundant signals (one for each channel) that can be used to enhance the processed signal, even if some of the channels have a very poor signal to noise ratio; and b) there is information encoded in the signals about the spatial position of the source (speaker) that is different from one to another. In previous work [11], a processing technique using the time delay of arrival (TDOA) was applied to the different microphone channels by delaying in time and summing the channels to create an enhanced signal. With this enhanced signal, the speaker diarization error (DER) was improved by 3.3% relative compared to the single channel error for the RT05s evaluation set, 23% relative for the RT04s development set, and 2.3% relative for the RT04s evaluation set (see [12]for more information about the databases and the task).

Very few work, however has tried to use exclusively information on the location of the speakers to do speaker diarization. The only one that we are aware of is the work of Ellis and Liu [10]. They used the cross correlation between channels to find a peak that represents a delay value between two channels. They later clustered the delay values to create homogeneuous segments in the speech frames. The results they reported for the set of shows corresponding to

the RT04s development set is 62.3% DER* error.[1]   Other attempts to do just speaker segmentation (not clustering) based in location have been done by Lathoud et al [17]. They assume that the speaker is confined to a physical region and calculate time delays from the generalized cross-correlation between paired microphones within an array. The estimated time delays form input features that are integrated in a GMM/HMM framework to segment the audio. They also propose an extension to the method to handle the case of overlapping speech from multiple simultaneous speakers. Their results show that the segmentation using delay features is better than using LPCC features. However they use a database not standard, limited to 4 speakers, each talking 5 minutes and the test set is an artificial random combination of segments from 5 to 20 seconds in duration. It also used prior information about the number and location of speakers. A version that removes the assumption that the locations of the speakers are fixed and known a priori is given in [16].

   Furthermore, very few people has tried to merge location information with acoustic information in order to improve speaker diarization. One of the works published is by Ajmera, Lathoud and Cowan [15]. They demonstrate that the fusion of those two types of information improves speaker diarization. However, they used again a very specific database and specific evaluation criteria (not the DER) and mixed distant microphones with lapel microphones, so it is difficult to compare to other approaches.

   In the first part of this paper, we present some experiments to determine to what extent the time delay of arrival (TDOA´s) by themselves can be used to segment and cluster the different speakers in a room. We have tried to develop a system that is robust to the changes in the meeting conditions, room, microphones, speakers, etc. We present a method to use only the

---

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

delays to obtain a segmentation and cluster hypothesis. Using our method, we obtain a diarization error (DER*) [12] of 35.73% for the same set of shows that were used by Ellis and Liu (i.e. 42.64 % relative improvement). We show also the results obtained with other sets of shows as 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 useful for this task. By merging TDOA features and acoustic features we have been able to improve baseline results by 16.35% relative for the RT05s evaluation set 21% relative for the Devel06s database (see the explanation of the database content below) and 15% relative for the RT06s evaluation set. The performance of this method has matched the best system performance in the RT06s evaluation campaign [12].

The paper is organized as follows: In Section 2 we describe the basics of our system, in Section 3 we present the database used and the evaluation metric. Section 4 explains the basic diarization system using only acoustic data and multiple microphones. In Section 5 we introduce the first novelty of this paper, that is the use only of interchannels differences to do speaker diarization with a very good performance. Section 6 explains the technique that we have used to combine acoustic information and delay information to improve the performance of previous systems at ICSI. In section 7 we discuss the results obtained and the adavantages and drawbacks of our proposal. Section 8 is the conclusion.

II. SYSTEM DESCRIPTION
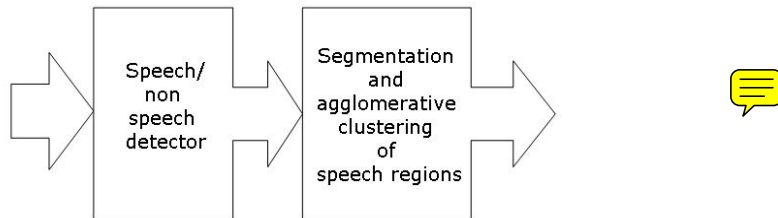
A. System architecture

Fig. 1. Speaker diarization system architecture

The general system architecture is presented in Fig.1. Firstly the utterance is segmented in speech/non speech regions. Then, the speech is segmented into homogeneous chunks, clustered and merged, and resegmented iteratively until a stopping criterion is reached.

B. Speech/non speech detector

One of the most important tasks in the process of speaker diarization is the separation of speech from all the other audio components, including silence, background noise, non-speech sounds as laughs, coughs, breaths etc. In fact the task is so important that NIST has decided to evaluate it also separately from the speaker diarization. The task is known as SAD (Speech Activity Detection). The detection of speech has a crucial role both in speaker diarization and speech recognition, because errors made at this stage cannot be recovered later. Because of the method used to calculate the Diarization Error (DER) every speech detection error is carried on

at the end of the process either as a False Alarm Error (speech is detected and there is no true speech in the reference) or a Missed Speech Error (non-speech is detected and there is speech in the reference)[2].

The question arises about what is the better method to do this task. Clearly the method is very much dependent on the application. It is also important to know also if there are training data available for the application or not.

Two methods have been used in this work depending on their availability at ICSI. The first method is provided by SRI and is based on a two class HMM decoder with a minimum duration of 30ms (three frames) enforced with a three-state HMM structure trained with Telephone Conversations and further tunned to RT02s data. The features used in the SNS detector (MFCC12) are different from the features used in the posterior process of segmentation and clustering. The resulting speech segments are merged to bridge short non-speech regions and padded according to NIST scoring guidelines. The speech/non-speech detector used is the same used in the RT05s evaluation campaign. The parameters of the detector were tuned on the RT05s meetings development data to minimize the combination of Misses and False Alarms as reported by the NIST mdeval scoring tool.

The second method has been developed by Xavier Anguera at ICSI and it does not need any previous data to train the system (although its parameters have been tuned with devel06 data). The method is based in an iterative two class segmenter that is started by initially considering non-speech all the frames that fall below a certain relative energy threshold. More information about it can be found in [21]. Although the method by itself produces worse SAD errors, the total DER is improved by using it. The method assumes that most non-speech segments are silence

---

[2] If there are more than one speaker in the reference, every Missed Speech Error will be multiplied by the number of speakers

(with low background noise) or close to it.
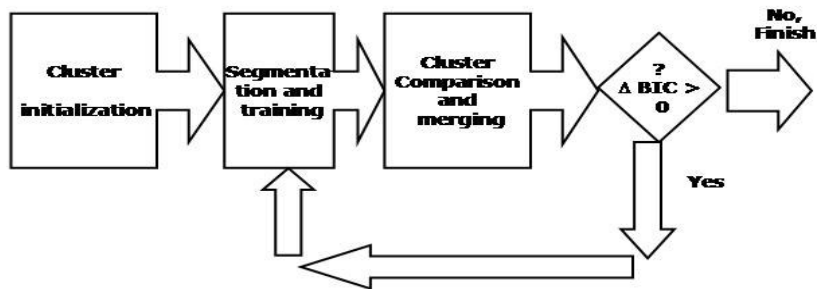
C. Iterative segmentation and clustering

Fig. 2. Segmentation and clustering process

The segmentation and agglomerative clustering process was proposed originally by Ajmera et al [4] and is shown in Fig. 2. The first module is the initialization, which will be explained later. Then a segmentation is made (with Viterbi decoding using the initial HMM models) and a training step. This process may iterate several times. Then a cluster comparison and merging is done. When a merging is done, the GMM for a new cluster is retrained with the new data assigned to it. Then it follows a subsequent segmentation and training and a new merging until a stopping criterion is reached. The system consists of an ergodic HMM with a number of states equal to the number of speakers (clusters) (see Fig. 3). The cluster model consists of a set of substates, the number of substates determined by the minimum duration in frames attributed to each cluster. Every state is modeled with a Gaussian Mixture of a number of components that has

to be set up initially. After the minimum number of substates, the system can jump to a new cluster or stay in the same. The jumping is determined in our case only by the acoustics (no probability is applied to the last state, alpha and beta equals one) see Fig. 4.
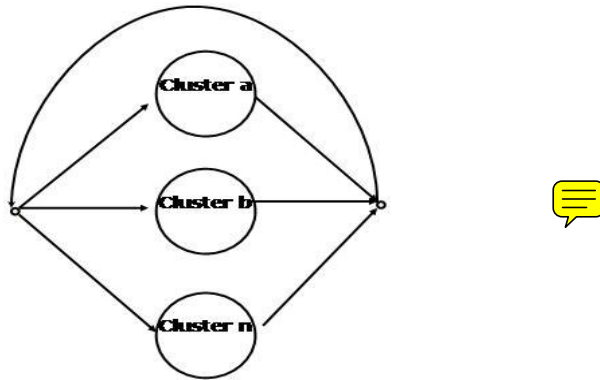

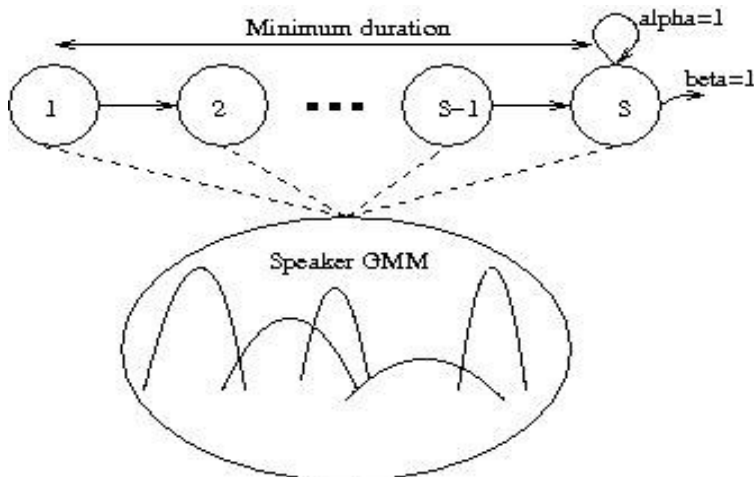Fig. 3 . Model for the system (After Ajmera and Wooters)

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

D. 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, giving good results [3],[14] and the modification of ΔBIC to eliminate the need of a penalty term has also given us good results [4]. The modified ΔBIC that we use for merging clusters is:

$$\Delta BIC = \log \ p(D|\theta) - \\ \log \ p(D_a|\theta_a) - \log \ p(D_b|\theta_b) \quad \textbf{(1)}$$

θ_a is the cluster created from D_a and θ_b is the cluster created from D_b, θ is the model created from D which is 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 that are used in the comparisons, particularly if the number of parameters in the merged cluster has to be the addition of the number of parameters of the merging clusters (see [22]). If ΔBIC is greater than 0 for a particular set of clusters, those two clusters are merged and the segmentation and training process is applied again to the new set of clusters. We merge the clusters that get the greatest ΔBIC. This basic method has been applied at ICSI since 2003 to do speaker diarization with a single channel source (single microphone).

E. Initialization

Ajmera showed in his paper that the initialization procedure was not relevant in the process. He segmented the data into K parts equally divided. This data was used to train the initial GMM models. An additional loop of segmentation and training can be done before going to the next

module. Other experiments tell us that the initialization may be crucial in this process[22]. In this paper we will use the equally divided segmentation just mentioned.

III. DATA USED AND EVALUATION METRIC

A. Databases

In this paper we will used data coming from all the releases done by NIST concerning this task, in the years 2002-2006, RT02s, RT04s, RT05s and RT06s. When the experiments are done with a subset of the data, it will be specified. Particularly used has been a selection of data coming from RT02s, RT04s and RT05s. This selection has been 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 name this selection Devel06s set. The set is presented in Table 1.

TABLE 1: SET OF SHOWS IN DEVEL06S

AMI_20041210-1052
AMI_20050204-1206
CMU_20050228-1615
CMU_20050301-1415
ICSI_20000807-1000
ICSI_20010208-1430
LDC_20011116-1400
LDC_20011116-1500
NIST_20030623-1409
NIST_20030925-1517
VT_20050304-1300
VT_20050318-1430

B. Evaluation metric

The speaker diarization performance is evaluated comparing the hypothesis segmentation, given by the system, with the reference segmentation provided by NIST[12]. This reference segmentation was generated by hand according to a set of rules also defined by NIST. In the evaluation plan is also defined the evaluation metric and a program to calculate it from both transcriptions. The error obtained is called diarization error rate (DER) and it takes three errors

into account (Miss, FA and Spkr). The error is frame based. A Miss error occurs when a speech segment (possible overlapping speakers) is classified as non-speech. If there is more than one speaker talking, the Miss error is multiplied by the number of speakers. A FA error occurs when the system produces a speaker hypothesis when there is no speech in the reference. To calculate the Spkr error, the program ~~maps optimally~~ the hypothesis speakers to the reference speakers (only one reference speaker to one hypothesis speaker) so the overlap in duration between all pairs of reference and hypothesis speakers is maximum. The algorithm counts an error if a frame in the hypothesis is mapped to a wrong speaker in the reference.

Because the metric is time-based, it is weighted towards the loquacious speakers. An error for a speaker that do not speak much is less important that an error for a loquacious speaker, consequently, the DER obtained for a specific show may be very much dependent on how many speakers talk and the time relationship between loquacious and non loquacious speakers.

IV. Speaker diarization using acoustic features

The signals coming from the different microphones are Wiener filtered to improve the SNR as used in previous systems [23]. Then one of the signals (microphones), the one with lower SNR is selected as a reference channel. The time-delay between the other channels and the reference channel are calculated.

A. Time-Delays Calculation

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

We define the time delay of arrival (TDOA) of $x_i[n]$ with respect to $x_j[n]$, $d(i,j)$ as the time

difference of the sound arriving at each microphone.

In order to estimate the TDOA between segments, we used a modified version of the Generalized Cross Correlation with phase transform ($GCC_{PHAT}(f)$) [13] and estimate the delays between microphones with the following formula:

$$d(i, j) = \text{arg max}_d (R_{PHAT}(d)) \qquad (2)$$

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

B. Acoustic Fusion

Once the delays are calculated every 500 ms (with a window shift of 250 ms), the signals are delayed and summed together with a triangular window to generate a new composed signal (beamformed signal). The composed signal is then processed as a single signal. Mel Frequency Cepstral Coefficients (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). For the RT05s set of shows using this procedure we obtained 18.48 % DER (using the standard NIST scoring software without counting overlaps). It is worth to mention 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.

V. SPEAKER DIARIZATION USING ONLY BETWEEN CHANNEL DIFFERENCES

A. Baseline

For the speech regions we calculate the TDOA´s using the procedure mentioned in the previous section. This time the window shift is 10 ms, the same than the one used for acoustic features calculation. We form a vector of delays that has as many components as the number of

microphones minus 1. Non-speech frames are estimated previously and are excluded from the subsequent process. The vector of delays is then fed to the segmentation and agglomerative clustering module explained above substituting the acoustic vectors. We experimented with several set of parameters for the segmentation and agglomerative clustering. In contrast to the unsensitivity to the parameters when using acoustic vectors as mentioned by Ajmera [4], there is sensitivity to the parameters when we used only delay vectors. In Table 2 we show the DER for different set of parameters for RT05s set of shows using a minimum duration of 2 s.

TABLE 2. SPEAKER DIARIZATION ERRORS DER FOR THE RT05S MDM CONFERENCE ROOM EVAL SET

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

In the subsequent experiments we have used 1 initial mixture and 10 initial clusters. Obviously, if the number of speakers in the room is more than10, the errors of the system will dramatically increase. In Table 3 we present the DER for RT05s and the components of it (Miss Error, False Alarm Error, Sp/nSp and Spkr Error). Note that the Sp/nSp Error is the addition of the Miss Error plus the False Alarm Error.

TABLE 3. MISSED SPEECH, FALSE ALARM SPEECH, SPEECH/NON-SPEECH ERROR SPEAKER ERROR AND DIARIZATION ERROR FOR THE RT05S EVAL SET USING 1 MIXTURE AND 10 INITIAL CLUSTERS

File	Miss	FA	Sp/NSp	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

TABLE 4. DIARIZATION ERROR FOR THE RT05S EVAL SET COMPARING TDOA AND ACOUSTICS

File	TDOA	Acoustics
AMI_20041210-1052	16.53	9.5
AMI_20050204-1206	23.03	12.47
CMU_20050228-1615	18.28	8.37
CMU_20050301-1415	45.88	12.41
ICSI_20010531-1030	20.59	13.26
ICSI_20011113-1100	45.52	50.86
NIST_20050412-1303	25.19	13.42
NIST_20050427-0939	37.18	9.92
VT_20050304-1300	25.7	7.25
VT_20050318-1430	61.27	60.06
All	31.2	18.48

In Table 4 and Fig.5 we show DER for every show comparing the results using only the acoustics and the results using only the TDOA´s. The results using only the delays are more unstable and their variance across shows is bigger than the results using the acoustics. However, there are a set of shows where the results for every method are very similar. One can also notice that the results for a pair of shows is extremely bad. This effect may be due to several factors like total number of speakers and total number of turns.Some analysis of this behaviour, where some shows are difficult to analyze -nuts- and some others are very disperse comparing results obtained from different algorithms -flakes- applied to Broadcast News data can be seen in [24]. The authors conclude that the shows that are more difficult to diarize are the ones that contain a lot of speakers, a lot of speaker turns and the unexistence of a dominant speaker in time.
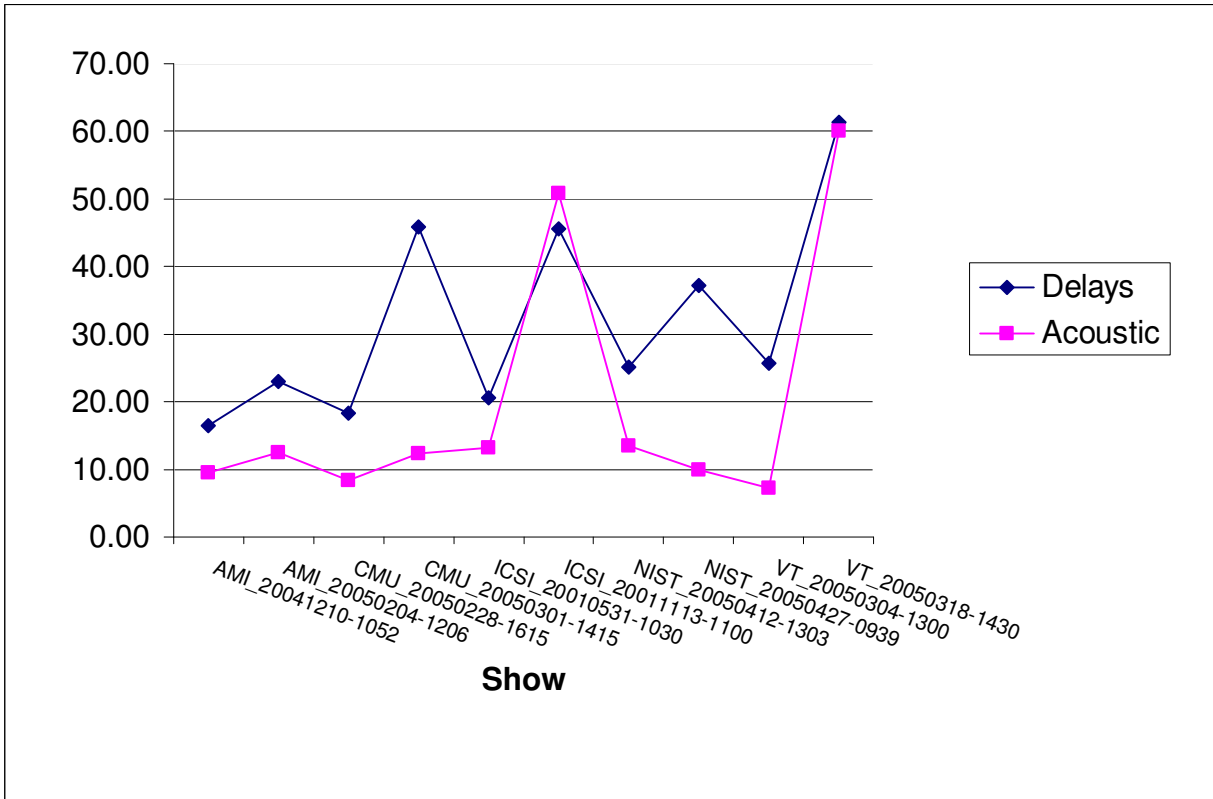
Fig. 5: Results across RT05s shows for two different systems: using acoustics only or delays only

TABLE 5. COMPARISONS BETWEEN RESULTS OBTAINED BY ELLIS AND LIU AND OUR RESULTS IN THE SAME SUBSET OF SHOWS FROM NIST RT04 DEVELOPMENT DATA- DER*.

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**%	

In order to be able to compare our results with the ones presented by Ellis and Liu [10], we have run the system with the same set of shows that they used in their experiments, and have reduced the number of channels available to 4 in all cases (Ellis and Liu used only 4 channels). In Table 5, the comparisons of both experiments are presented. It is important to notice that in these results, two of the shows from NIST RT04s development data (the CMU shows) have not been

used because the conditions of these shows (only one distant microphone) are not compatible with the conditions of our experiment (multiple distant microphones)[3]. Also the results presented here include the overlap regions and no False Alarms (we call it the DER[*] error). We have included in Table 5 also the standard DER error (including False Alarms) for completeness. The analysis of the results show a big improvement of our system compared to the Ellis one. The differences may well come from the different way we use to calculate the delays between signals and the different segmentation and clustering procedure. Since the number of microphones used in this experiment was less than the number of microphones available, we have also investigated the error rate that we could obtain for the same set of shows if we had used all the available microphones. Table 6 gives results of this comparison. It can be seen that the use of more microphones reduces the DER error rate by 3.26% absolute.

TABLE 6. COMPARISONS BETWEEN DER OBTAINED USING 4 CHANNELS AND RESULTS USING ALL THE CHANNELS AVAILABLE IN THE SYSTEM.

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

B. Delay calculation improvements

The calculus of the delays is not an exact procedure and sometimes due to several reasons, the autocorrelation between signals does not show the highest peak in the desired delay value. In this

---

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

paper we have also experimented with a new method to make more robust the calculation of the delays [25]. It consists of two phases. The first phase process N peaks (experiments done with 8) of the cross-correlation between the channel and the reference and makes a Viterbi alignment to extract the 2 best paths (2 best peaks) for every frame. For the Viterbi process, the emission probability used is the cross-correlation value and the transition probability between two nodes is the inverse of the difference between delay values ensuring that the N transition probabilities in a particular instant sum up to 1.The second phase process the best 2 peaks for every frame and every channel and makes a new Viterbi alignment between all channels to find the best path of the delay vector across all the sentence. The emission probabilities are the product of the individual correlation values of each considered delay and the transition probabilities are computed summing all delay distances from all considered delays and normalized to sum to 1. This technique aims at finding the optimum tradeoff between reliability (value of the cross correlation) and stability (distance between contiguous delays).

The DER for the Devel06s set using delays obtained by the baseline system is 35.39% and the DER obtained using the improved system is 29.45%. Thus for this set of shows we can see the improvement that we obtained by using the improved method. In contrast to the results presented up to now, the DER presented here has been evaluated using overlap regions and using also forced aligned transcriptions instead of the official ones made by hand.[4] The forced aligned references 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 [8]. Note that the baseline value differs from the one presented later in Table 8 due to a change in parameters of the speech/non speech detector used.

VI. 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 try to combine them. The first idea that we had is to concatenate both vectors, 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. So we decided to keep both vectors separately and model the clusters with independent information coming from both set of vectors. The general architecture of the system is depicted in Fig. 6.
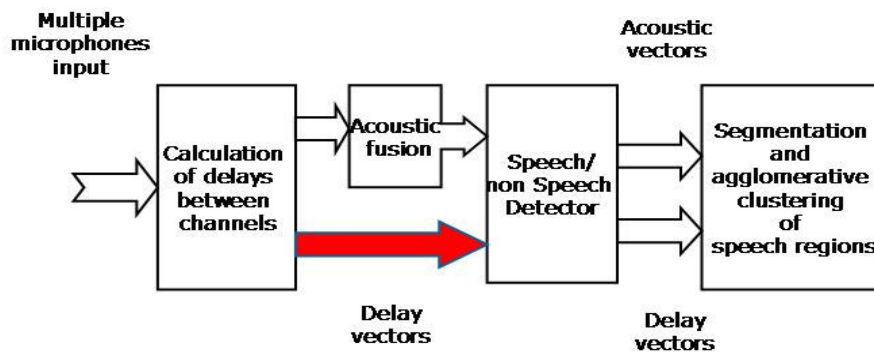
Fig. 6: General architecture of the system that uses both acoustic vectors and delay vectors.

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 from non speech. The speech regions are then processed to obtain the MFCC as mentioned in

---

[4] The reason for it is that we had several concerns about the validity of the official transcriptions to compare and improve systems, because of the method used (using head mounted microphones) and the problems with labelling the non-speech regions.

section IV. 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 also the clusters to merge which needs also the computation of the log likelihood of a set of vectors given the model. For the combined system we used a joint log likelihood as following:

$$\log \ p(x[n],\ y[n]\ |\ \theta_a) =$$
$$\alpha \log \ p(x[n]\ |\ \theta_{ax}) + (1 - \alpha) \log \ p(y[n]\ |\ \theta_{ay}) \qquad (3)$$

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] associated to it and θ_{ay} is the model created for cluster a using the delay vectors y[n] associated to it. α is a weight factor that has to be determined. The different DER for RT05s without overlap as a function of the weight factor applied is presented in Fig 7. Starting from values of 31.2% for delays only and 18.48% for acoustic only, we obtain 15.46 for the compound system for a weight of 0.9 (using a minimum duration of 2s and an initial number of clusters of 10). This implies a reduction of DER of 16.34% relative. Equally 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 forced aligned labels have been used in scoring. Also for this experiment the speech/non speech detector has been changed and the one in [21] has been used. This detector does not need any training data. From 31.97% using only delays and 12.71% using only acoustic features we obtain a DER of 10.04% with a weigth factor also of 0.9. This implies a reduction of DER of 21%. In Table 7 we give details of the DER obtained for each show, for future reference.

In Table 8, in the first two columns we present the data mentioned above. In the third column

we present the results obtained at the official RT06s evaluation campaign (35.77% DER which obtained the best score). For the evaluation campaign, the only change that we did comparing to the Devel06s system is the use of the improved delay calculations presented in Section 5 B. We also show the results with the error obtained after the evaluation with acoustic only data and delays only data, getting a 15.1 % relative improvement over the acoustic only result.

In Table 8 we also present the data scored with forced aligned labels (4th column). The relative improvement obtained using forced aligned labels is bigger (25.84%).
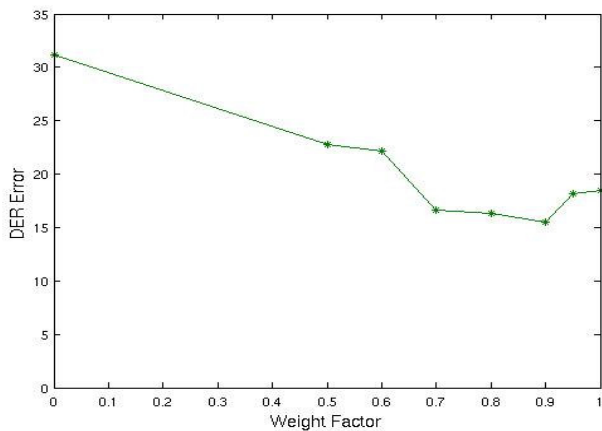
Fig.7 : Plot of different DER errors as a function of the weight factor applied for the RT05s
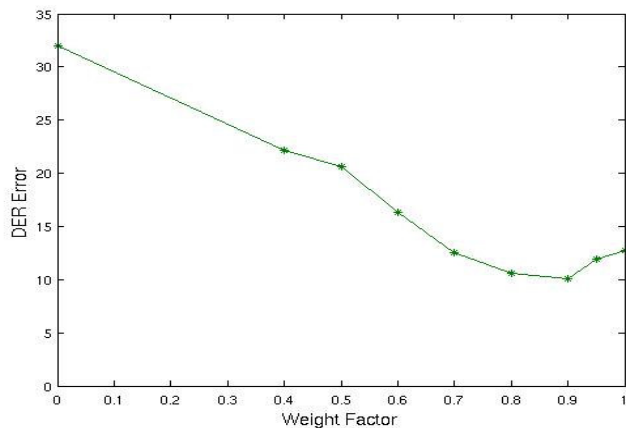
Fig. 8: DER as a function of the weight factor for the Devel06s data

TABLE 7: RESULTS FOR THE SET DEVEL06S. WE PRESENT THE PERCENTAGE OF MISSED SPEECH, FALSE ALARM SPEECH,, SPEAKER ERROR AND TOTAL DIARIZATION ERROR (DER).

```
      File        Miss   FA    Spkr   Total

 AMI_20041210-1052  0.40  1.20   1.10    2.69
 AMI_20050204-1206  2.60  2.20   3.30    8.01
 CMU_20050228-1615  9.30  1.20   1.80   12.30
 CMU_20050301-1415  3.70  1.60   1.10    6.41
ICSI_20000807-1000  4.60  0.40   3.80    8.77
ICSI_20010208-1430  3.60  1.10  11.00   15.72
 LDC_20011116-1400  2.10  3.00   4.20    9.32
 LDC_20011116-1500  5.90  1.10   7.60   14.65
NIST_20030623-1409  1.00  0.70   1.40    3.08
NIST_20030925-1517  7.70  5.70   9.60   22.95
  VT_20050304-1300  0.60  1.00   2.80    4.43
  VT_20050318-1430  1.30  6.20  13.80   21.36
               ALL  3.40  1.90   4.70   10.04
```

TABLE 8: DER ERROR FOR THE EVAL05S AND DEVEL06S AND THE OFFICIAL RT06S DATASET OBTAINED USING ACOUSTIC FEATURES ONLY, DELAY FEATURES ONLY AND COMBINED FEATURES

Features used	RT05s (hand labels-no overlap)	Devel06s (forced aligned labels-overlap)	RT06s (hand labels-overlap)	RT06s (forced aligned labels-overlap)
Delays only	31.20 %	31.97 %	47.55% (after the eval)	
Acoustic features only	18.48 %	12.71 %	42.13% (after the eval)	27.01% (after the eval)
Combined acoustic+delays	15.46 %	10.04 %	**35.77%** (official result)	20.03% (pseudo-official)
Relative error reduction	16.34 %	21 %	15.10% (after the eval)	25.84% (after the eval)

.

If we examine Table 8, we can notice that the relative improvement of the combined system compared to the acoustic only system is bigger when we use force aligned labels (even if the base error is less) 25.84% versus 15.10%. One question arisen in the last evaluation campaign was the appropriateness of using hand labels to compare and evaluate systems specially when overlapped speech is included in the evaluation. During this year´s development period we experienced difficulties when using had-made reference files, mostly when scoring on speaker overlap 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 overlap labels on non-speaker-overlap regions (because the hand references have been created with close microphones, the overlap may be noticed on the IHM channels but its volume is too low to be perceived by the MDM channels). All these artifacts create an extra amount of missed-speech error and of speaker error which is not consistent over the different evaluation sets. Consequently, for this year´s system development we have taken the initiative to use references derived from forced alignments (results showed in the second column of Table 8).

VII. DISCUSSION AND IMPROVEMENTS

We have mentioned the problem of DER for different shows in Section 5 in which some shows give very good DER results and some others give much worse results. The method that we have used to segment the speakers using only delays assumes that the speakers are not moving very much from their position. In other words, the system assigns a speaker to a region in space. If this is not the case, the DER using TDOAs only will be severely increased. Alternatively, using only acoustics, some shows give very bad performance, one of the reasons maybe the existence of speakers very close in the acoustic dimension. The interest of mixing delay information and acoustic information delivers a robust system against weakness in either one or the other dimension.

We have used a weight factor between delays and acoustics fixed and optimized according to a development set. But again the optimal weight factor may be dependent on the show itself. To illustrate this problem, in Fig. 9 we show the DER across different weights for the Devel06 set (labelled Average all) against the same plot for a subset of it (labelled Average subset) which

corresponds to the shows CMU_20050228-1615, LDC_20011116-1500 and VT_20050304-1300). We can observe that the minimum value for the subset of shows appears in the weight factor 0,7 that may well correspond to the case of several speakers very close acoustically but located in clearly separated areas of the room and staying still.
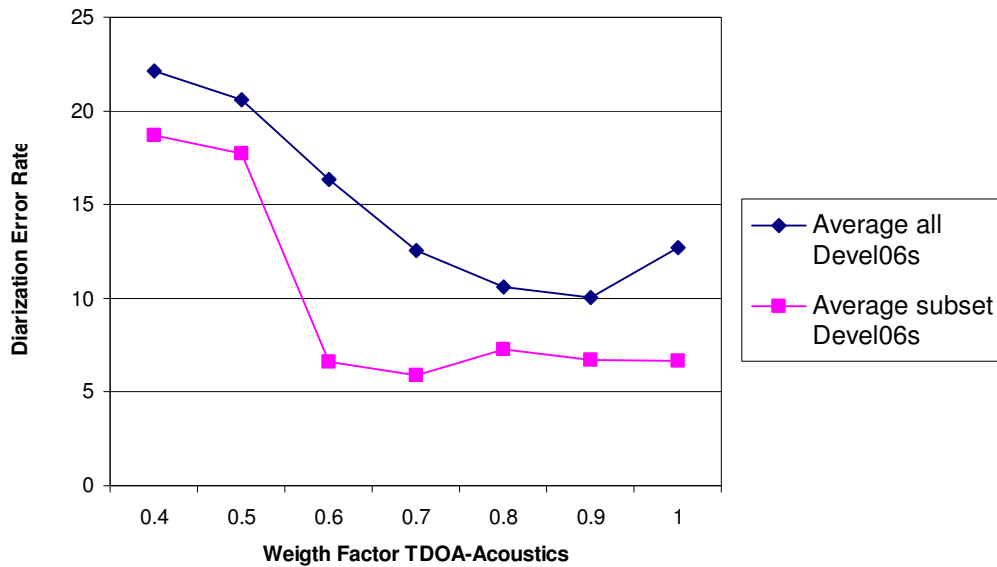

Fig 9: Comparison of DER across different weight factors for the develo06s set (Average all) and for a subset of it (Average subset)

The results presented here could be further improved by improving separately the discrimination capability of both methods. By using the improved method of calculating the delays presented in section 5 B, we have been able to get better results for the Devel06s set and decrease the DER to 9.74% absolute (3.4% relative) (with overlap and force aligned references). Those results could be further improved if we mixed other sources of information as pitch, as preliminary experiments done in our laboratory show [26].

VIII. CONCLUSION

We have proposed and develop 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 in the path of combining as many sources of information as possible to solve the problem. Particularly interesting could be the inclusion of suprasegmental information, such as pitch, language models etc. and other techniques applied to speaker verification systems. An important area of research is to develop 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, so to make discriminative and interactive communication between humans and computers in meetings an attainable goal.

ACKNOWLEDGMENT

REFERENCES

[1] http://nist.gov/speech/tests/rt/rt2002/
[2] D. A. Reynolds, P. Torres-Carrasquillo : Approaches and applications of audio diarization, *ICASSP 2005*, pp, V-953-V 956, 2005.
[3] J. Ferreiros, D. Ellis: Using acoustic condition clustering to improve acoustic change detection on broadcast news. *Proc. ICSLP 2000*
[4] J. Ajmera, C,Wooters : A Robust speaker clustering algorithm, *IEEE ASRU 2003*.
[5] X. Anguera, C. Wooters, B. Pesking and Mateu Aguiló : Robust speaker segmentation for meetings: the ICSI-SRI spring 2005 diarization system, *Proc NIST MLMI Meeting Recognition Workshop*, Edinburgh, 2005
[6] 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" *Proceedings of the Joint AMI/Pascal/IM2/M4 Workshop on Meeting Recognition*. Also published in *Lecture Notes in Computer Science*, Volume 3361 / 2005.
[7] C.Wooters, J, Fung, B. Pesking, X. Anguera, "Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system" *NIST RT-04F Workshop*, Nov. 2004.
[8] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters and J. Zheng, "Further progress in meeting recognition: the ICSI-SRI spring 2005 speech-to-text evaluation system" *Proceedings of NIST MLMI Meeting Recognition Workshop*, Edinburgh.
[9] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters and B. Wrede, "The ICSI meeting project: resources and research" *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal
[10] D.P.W Elis and Jerry C.Liu : Speaker turn segmentation based on between-channels differences, *Proc. ICASSP 2004*.
[11] X. Anguera, C. Wooters, J. Hernando : Speaker diarization for multi-party meetings using acoustic fusion, *IEEE ASRU, 2005*.

[12] NIST Spring 2006 (RT06S) Rich Transcription Meeting Recognition http://www.nist.gov/speech/tests/rt/rt2006/spring/

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

[14] S.S. Chen, P.S. Gopalakrishnan: speaker environment and channel change detection and clustering via the bayesian information criterion, *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998

[15] J. Ajmera, G. Lathoud, I..A.Mc Cowan : Clustering and segmenting speakers and their locations in meetings, *ICASSP 2004*, pp I-605, I608, 2004.

[16] G. Lathoud, I..A.Mc Cowan, J.M.Odobez : Unsupervised location-based segmentation of multi-party speech, *ICASSP-NIST Meeting Recognition Workshop*, Montreal, Canada May 2004.

[17] G. Lathoud, I..A.Mc Cowan : Location based speaker segmentation, *ICASSP 2003*, I-176, I-179, 2003.

[18] S.E. Tranter, D. A. Reynolds, Speaker diarisation for broadcast news, *Proc Odyssey: The Speaker and Language Recognition Workshop*, p 337-344, Toledo, Spain, May 2004.

[19] C. Barras, X.Zhu, S. Meignier and J.L. Gauvain, Improving speaker diarization, *Proc. DARPA RT04*, Palisades NY, November 2004

[20] J.M.Pardo, X. Anguera, C. Wooters: Speaker diarization for multi-microphone meetings using only between-channel differences. *Proc. MLMI 06, 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 1-3 May 2006, Washington DC, USA. To appear in *Lecture Notes in Computer Science*.

[21] X. Anguera, M Aguiló, C. Wooters, C. Nadeu, J. Hernando "Hybrid speech/non-speech detector applied to speaker diarization of meetings" *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop* 28-30 June 2006 ,San Juan, Puerto Rico

[22] X. Anguera, C. Wooters and J. Hernando, "Automatic cluster complexity and quantity selection", *in MLMI´06,* Washington DC, USA, Msy 2006.

[23] N. Mirghafori et al. "From switchboard to meetings: Development of the 2004 ICSI-SRI-UW meeting recognition system", *ICSLP 04*, Korea, Ocotber 2004.

[24] N. Mirghafori and C. Wooters, "Nuts and flakes: A study of data characteristics in speaker diarization", *ICASSP06.*

[25] X. Anguera, C. Wooters, J. M. Pardo" Robust speaker diarization for meetings: ICSI RT06s meetings evaluation system" To appear in *Lecture Notes in Computer Science*.

[26] A. Gallardo-Antolín, X. Anguera, C. Wooters, "Multi-Stream Speaker Diarization Systems for the Meetings Domain", in ICSLP 06, Pittsburg, September 2006.

[27] Sylvain Meignier, Daniel Moraru , Corinne Fredouillea, Jean-François Bonastre, and Laurent Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization " *Computer Speech and Language*, vol 20 pp 303-330, April-July 2006

[28] S. Tranter, D. A. Reynolds, "An overview of automatic speaker diarization". *IEEE Trans. On Audio, Speech and Language Engineering, Vol 14, No.5*, pp 1557- 1565, September 2006

[29] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, L. Bonastre, "The ELISA consortium approaches in broadcast nes speaker segmentation during the NIST 2003 rich transcription evaluation", *ICASSP 2004*, Montreal, Canada

[30] X. Zhu, C. Barras, S. Meignier, J.L. Gauvain, "Combining speaker identification and BIC for speaker diarization in *Proc.Eur. Conf. Speech Commun. Tehcnol.* Lisbon, Sep 2005, pp 2441-2444.

[31] D. A. Reynolds and P. Torres -Carrasquillo, *"The MIT Lincoln Laboratory RT04F diarization systems: Applicatins to broadcast audio and telephone conversations, in Proc. Fall Rich Transcription Workshop (RT-04),* Palisades, NY, 2004.

[32] R. Sinha, S. E Tranter, M.J.F. Gales and P.C. Woodland, "The Cambridge University March 2005 speaker diarization system, in *Proc.Eur. Conf. Speech Commun. Tehcnol.* Lisbon, Sep 2005, pp 2437-2440.

Jose M. Pardo (M'84-SM'04) got a Telecommunication Engineering Degree and a PhD both from Universidad Politécnica de Madrid in 1978 and 1981 respectively. He got 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 held different teaching and research positions at the Universidad Politécnica de Madrid. He is Full Professor since 1992 and Head of Speech Technology Group since 1987. He was head of the Electronic Engineering Department from 1995-2004. Prof Pardo has been a Fulbright Scholar at MIT in 1983-84 ,a visiting scientist at SRI International in 1986 and recently at the International Computer Science Institute in 2005-2006. He is member of JASA, ISCA and EURASIP. He is member of the ISCA Advisory Council. He was chairman of EUROSPEECH 1995 and member of ELSNET Executive Board 1998- 2004 and member of NATO RSG 10 and IST 3 since 1994. Since 1978 he works in Speech Technology. He has more than 160 papers and holds two patents.

Second A. Author (M'76-SM'81-F'87) and the other authors may include biographies at the end of regular papers. Biographies are not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. Other usual biography information includes birth date and place, education, employments, and memberships in other professional societies.

Third A. Author (M'76-SM'81-F'87) and the other authors may include biographies at the end of regular papers. Biographies are not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. Other usual biography information includes birth date and place, education, employments,   and   memberships   in   other   professional   societies.