Speaker Diarization Features: The UPM Contribution to the RT09 Evaluation

José M. Pardo, *Senior Member, IEEE*, Roberto Barra-Chicote, Rubén San-Segundo, Ricardo de Córdoba, *Member, IEEE*, and Beatriz Martínez-González

***Abstract*—Two new features have been proposed and used in the Rich Transcription Evaluation 2009 by the Universidad Politécnica de Madrid, which outperform the results of the baseline system. One of the features is the intensity channel contribution, a feature related to the location of the speaker. The second feature is the logarithm of the interpolated fundamental frequency. It is the first time that both features are applied to the clustering stage of multiple distant microphone meetings diarization. It is shown that the inclusion of both features improves the baseline results by 15.36% and 16.71% relative to the development set and the RT 09 set, respectively. If we consider speaker errors only, the relative improvement is 23% and 32.83% on the development set and the RT09 set, respectively.**

***Index Terms*—Features for speaker diarization, speaker diarization, speaker segmentation, speech processing in meetings.**

I. Introduction

SPEAKER diarization is the task of identifying the number of participants in a meeting and creating a list of speech time intervals for each participant. Speaker diarization is useful as a first step in the speech transcription of meetings in which each spoken sentence has to be assigned to a defined speaker. It can also be used for speaker adaptation in speech recognition.

An overview of automatic speaker diarization systems is given in [1].

Common speaker diarization systems consist of three main blocks: the voice activity detection module (VAD), the feature extraction module and the segmentation and clustering module; see Fig. 1.

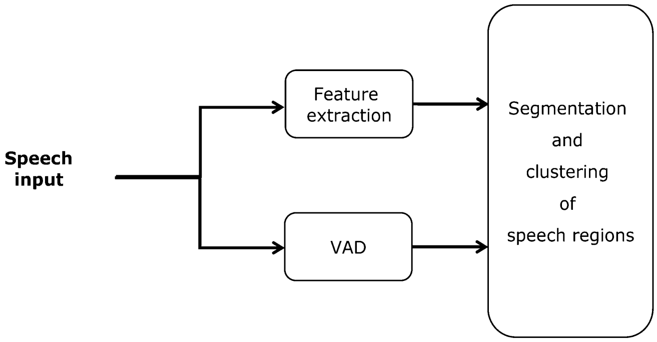

Fig. 1. Simplified diagram of a speaker diarization system.

VAD algorithms differ, depending on the type of non-speech sounds that appear next to the speech or mixed with it, from the Gaussian mixture models (GMMs) to Laplacian and gamma probability density functions [2]. Voice activity detection is, by itself, a large area of research. Voice activity algorithms applied to speaker diarization may differ from general algorithms because the diarization error rate is measured frame by frame instead of other metrics that ponder the error based on the correctly detected speech/silence segments.

The feature extraction module usually extracts data related to the spectral envelope such as the Mel frequency cepstral coefficients (MFCCs) [3], [4].

Regarding segmentation and clustering of speech regions, some methods use bottom-up agglomerative clustering [5], [6], while others use a top-down universal background model (UBM) as a starting point to apply adaptation techniques iteratively to build the speaker models [7]. Clustering algorithms need a distance measure to determine whether two speech clusters belong to the same speaker. The most common used distance is the Bayesian information criterion (BIC) distance [8]. Recent studies have also presented great improvements using other alternatives based on the t-test distance [9] or the information theoretic approach [3].

Speaker diarization was first applied to broadcast news recordings (BN). One of the best recently published systems can be seen in [10]. Subsequently speaker diarization was applied to the meeting domain using a single distant microphone (SDM). 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 as a first step toward the development of algorithms for MDM.

Manuscript received July 29, 2010; revised December 23, 2010 and April 01, 2011; accepted May 26, 2011. Date of current version January 13, 2012. This work was supported in parts by projects SD-TEAM: TIN2008-06856-C05-03/TIN from the Ministry of Science and Innovation, ROBONAUTA, DPI2007-66846-C02-02, and INAPRA DPI2010-21247-C02-02 from Ministry of Education and Science, and MA2VICMR, S2009/TIC-1542 by the Comunidad Autónoma de Madrid. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Friedland.

The authors are with the Speech Technology Group, Universidad Politecnica de Madrid, ETSI Telecomunicación, Ciudad Universitaria s/n, 28040 Madrid, Spain (e-mail: pardo@die.upm.es; barra@die.upm.es; lapiz@die.upm.es; cordoba@die.upm.es; beatrizmartinez@die.upm.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2159971

In speaker diarization with multiple distant microphones (MDMs) redundant information is available (one signal per microphone) in comparison with single distant microphone (SDM) diarization. Usually, all speech signals are combined into one [11], from which some acoustic features are extracted. The other source of information used in MDM scenarios is the information related to speaker localization [12], such as the time delay of arrival (TDOA) features [13]. 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. In [14], it was first demonstrated that TDOA between channels could be mixed with spectral features to obtain improved performance over a base system that used only spectral features. This TDOA information combined with the MFCC information has been used by all systems in the latest Rich Transcription evaluation [15].

The shortcomings of TDOA methods are the result of distant microphones. There are noises and reverberations in the recordings and the results are not free from errors. In speaker diarization in MDM scenarios, not only the improvement of the VAD module or the segmentation and pattern classification modules is necessary. It is also important to search for new features that convey additional information to improve system performance [16].

In [17], a method to improve inaccurate estimates of delays and increase speaker separation in delay-space was presented.

In [18], the logarithm of F0 plus its derivative were used successfully in a speaker diarization for single distant microphone meetings (SDMs) using a method to normalize the features across all speakers and combine them with Mel frequency coefficients (MFCC) at the segmentation phase and using MFCC features only at the clustering phase. In [16], the use of the F0 average and the median F0 calculated on a 500-ms Hamming window and several other so-called long term features to improve the performance of an MFCC-based system applied to SDM meetings were proposed. The authors point out the importance of long-term features (longer than a frame) in speaker discrimination and speaker diarization task. In [19], the authors have used long-term and prosodic features for clusters initialization for MDM meetings.

In previous work [20], we developed a method to combine MFCC coefficients with a time delay of arrival features (TDOA) to create an enhanced system for multiple distant microphone meetings (MDMs).

In this paper, we present two new features that improve speaker diarization for MDM meetings which were included in the last submission by the Universidad Politécnica de Madrid (UPM) to the National Institute of Standards Rich Transcription Evaluation in 2009 (NIST RT 2009).

The first feature is related to the localization of the speakers (similar to TDOA features) that we called the intensity channel contribution (ICC) and which makes use of the normalized energy of the signal arriving at the different channels [21]. It is the first time that such a feature is proposed and used in speaker diarization.

The second feature is based on the use of the fundamental frequency (F0) but instead of using it for SDM meetings as in [18], or [16] we have used it for MDM meetings. It is the first time that it has been used for MDM meetings in the segmentation and clustering stage. Instead of using it in the segmentation stage as in [18] we have used it both in the segmentation and agglomerative stages similar to [20]. In contrast to using it as a long-term feature (500-ms span) as in [16] we have used it as a frame-based feature (20 ms). We also present in the paper experiments using different methods to include F0 and different methods to combine them with MFCC, TDOA, and ICC features.

By using ICC features, we have been able to improve the baseline system DER by 4.6% and 7.9% relative for the development set and the RT09 set, respectively. By using F0 we have improved the baseline system DER by 7.31% and 10.63% relative for the development set and the RT09 set, respectively, and finally using both ICC and F0 we have improved DER by 15.3% and 16.7% for the aforementioned databases. A large part of the DER comes from the speech/non speech errors. If we take into account just the speaker errors, the improvement in the proposed system is 23.4% and 32.83% relative on the development set and the RT09 set, respectively.

Since the features module is very independent of other modules we think that the proposed system could contribute to the improvement of alternative state of the art systems.

The paper is organized as follows. In Section II, the baseline system is described. In Section III the proposed new features are presented. Section IV describes the corpora used both for development and test and describes the evaluation metric. Section V includes the experiments carried out and the results obtained. Section VI is the discussion of the results and finally Section VII ends with the conclusions.

II. Description of the Baseline System

A. Baseline System Architecture and Baseline Features

Fig. 2 shows the system architecture. The input coming from several different microphones (D_{ICC}) is first Wiener filtered in order to reduce the background noise.

Then, in order to estimate the TDOA between two segments from two microphones, we use a modified version of the Generalized Cross Correlation (GCC) called "generalized cross correlation with phase transform" (GCC-PHAT) [22]. First, one of the channels is selected as the reference channel (the one with highest cross correlation with other channels). Then the GCC-PHAT between the reference channel and the other channels is estimated and the TDOA for these two microphones is calculated as

$$\mathrm{TDOA} = d(i,j) = {}_{d}^{\arg\max}(R_{\mathrm{PHAT}}(d)). \tag{1}$$

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

The set of TDOAs from each microphone to the reference channel will form what we call the TDOA vector $\mathbf{tdoa}$ which has a $D_{\mathrm{ICC}} - 1$ dimension. Once the $\mathbf{tdoa}$ vector is calculated,

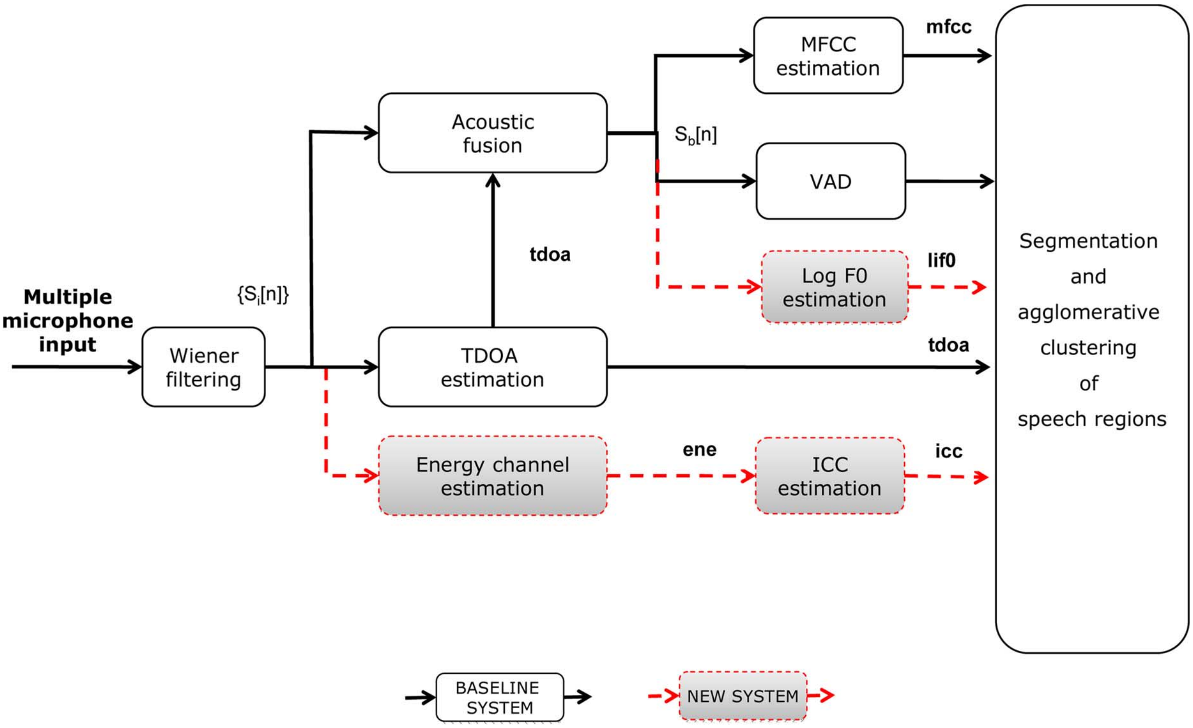

Fig. 2. Proposed system architecture.

a weighted delay-and-sum algorithm is applied in the acoustic fusion module, where the input signals are delayed and added together to generate a new composed signal. More details on this part can be found in [11]. The composed signal is then processed by the MFCC estimation module, where MFCC vectors of 19 components, **mfcc**, are calculated every 10 ms with a window of 30 ms.

The VAD module is a hybrid energy-based detector and model-based decoder. In the first stage, an energy-based detector finds all segments with low energy, while applying minimum segment duration. An energy threshold is set automatically to obtain enough non-speech segments. The segmentation is used to train speech and non-speech models in the second module and then several iterations of Viterbi segmentation and model retraining take place, finally outputting the speech/non-speech segmentation when the likelihood converges. More information on the VAD module can be found in [23].

The segmentation and agglomerative clustering process consists of an initialization part and an iterative segmentation and merging process [24]. The initialization process segments the speech into K blocks (equivalent to an initial hypothesis of K speakers or clusters) uniformly distributed. We have set K to 16 empirically.

An individual cluster model consists of a set of sub-states, where the number of sub-states is determined by the minimum duration of each cluster, 2.5 seconds in our case. Every sub-state is modeled using a Gaussian mixture model (GMM) initially containing a number of components that has to be specified (we use 5 for **mfcc** and 1 for **tdoa** streams). After the initial segmentation a set of training and re-segmenting steps is carried out using Viterbi decoding. Then the merging step takes place. 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 segmentation and clustering steps are repeated until a stopping criterion is reached; see Fig. 3.

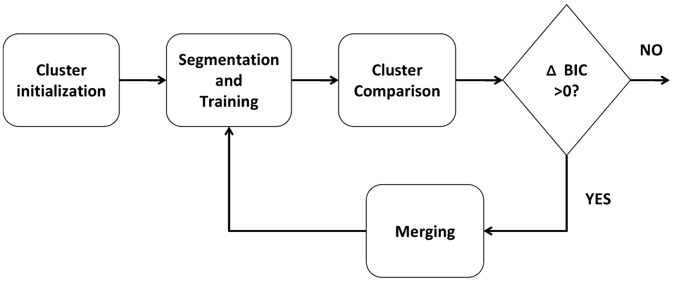

Fig. 3. Block diagram of the segmentation and clustering method.

To decide which clusters to merge, and when to stop the merging, the BIC criterion has been used. The penalty term λ in the BIC score is eliminated because we constrain both hypotheses to have the same number of parameters [24]. When all possible merge pairs give a negative BIC, the merging is stopped.

Some percentage of frames (silences, noises) constitute a different set and are too short to be part of a new cluster but corrupt the cluster models [25]. A frame purification algorithm is applied which aims to detect and eliminate non-speech frames that do not help in discriminating speakers. 10% of frames with the highest likelihood computed on Gaussians with smaller variance are removed for training models that have more than two Gaussians before computing the BIC.

The baseline features used in the diarization task are the MFCC features combined with the TDOA features. In the implemented system [20], the first 19 MFCC coefficients are extracted and treated as the $\mathbf{x}$ stream and the TDOA features are treated as the $\mathbf{y}$ stream. Each source of information is modeled using a statistical model whose individual likelihoods are combined using

$$\log p(\mathbf{x},\mathbf{y}|\theta_a) = w_x \log p(\mathbf{x}|\theta_{ax}) + w_y \log p(\mathbf{y}|\theta_{ay}) \qquad (2)$$

keeping $w_x + w_y = 1$. θ_a is the compound model for any given cluster a, θ_{ax} is the model created for cluster a using the stream $\mathbf{x}$, and θ_{ax} is the model created for cluster a using the stream $\mathbf{y}$. This baseline system is similar to the system presented by the International Computer Science Institute (ICSI) at the RT06 evaluation in which the first author of this paper was a team member, obtaining state of the art performance.[1]

III. Proposed New Features

A. ICC Features

The first contribution of UPM (Universidad Politécnica de Madrid) to the RT09 evaluation was the inclusion of a new set of features related to the localization of the speakers. Apart from the delay vector made up of the delays between every channel and the reference channel we have computed the relative energy for each channel and frame compared to the sum of the energies for all the channels

$$\mathrm{icc}[i] = \frac{\mathrm{ene}[i]}{\sum_{i=0}^{D_{\mathrm{ICC}}-1} \mathrm{ene}[i]} \qquad (3)$$

in which $\mathrm{icc[i]}$ is the intensity channel contribution per frame, i is the channel number, $\mathrm{ene[i]}$ is the energy per frame, and channel and D_{ICC} is the number of channels as we mentioned before. The vector of $\mathrm{icc[i]}$ for each frame is concatenated to the TDOA vector $\mathbf{tdoa}$ to form the $\mathbf{tdoa}+\mathbf{icc}$ vector. Note that the $\mathbf{tdoa}$ vector has $D_{\mathrm{ICC}} - 1$ components, one less than the ICC vector. The energy captured by each channel is related to the distance of the speaker to that particular microphone: when higher energy is detected, it means that the speaker is closer to that microphone. This is related to the localization of the speaker similar to the information conveyed by the TDOA features. The difference is that the signal delay information used in the estimation of the TDOA features is proportional to the distance, while the intensity is inversely proportional to it.

The consideration of both features, TDOA and the proposed energy related features, assumes that the speakers do not move around the room. Note also that although apparently both absolute energy $\mathrm{ene[i]}$ and ICC features $\mathrm{icc[i]}$ are obtained from the same measure (the energy) if the same speaker, at a certain location, augments his intensity level from one turn to another, the absolute energy features computed at each channel will have a bias corrupting the speaker models while the ICC features will not have this problem resulting in a more robust set of features. An exhaustive analysis of the behavior of energy features and ICC features and improvements that can be obtained by using them can be seen in [21].

B. F0 Features

At the RT09 evaluation we have successfully used F0 features to improve the diarization performance of MDM meetings. In order to determine the way of calculating F0, we have experimented with different parameters.

First, F0 is calculated using the algorithm in [27]. For each frame we take a window of about 7.5 ms and calculate its normalized cross-correlation with the speech signal in windows at various "lags" in the future. Lags range from less than 2 ms (for f0 = 500 Hz) to more than 20 ms (for f0 = 50 Hz).

Then the logarithm of F0 was calculated. For the unvoiced part of the signal a constant value of F0 was used which is the average of the last value of the previous voiced region and the first value of the following voiced region. We will call this feature **lif0** from now on. We have also experimented with the plain interpolated F0, called **if0** from now on. Similarly, a third feature was the first derivative of the logarithm of the interpolated F0 and called **dlif0** from now on.

Finally, a fourth method of calculating F0 has been researched using a long-term window. F0 is estimated frame by frame (10-ms frame shift). Then, a histogram of the F0 values is calculated using a window of 500 ms (50 frames). 23 bins are used: 19 bins (from 60 Hz to 250 Hz, with 10-Hz resolution), 3 bins (from 250 to 310 Hz, with 20-Hz resolution) and 1 bin for F0 values higher than 310 Hz. The counts of the histograms are normalized by the number of total observations (50 observations, equal to the number of frames) and used as a feature vector. This feature vector will be called **hf0** from now on.

C. Feature Combination

It is not trivial how to combine different features in speaker diarization since they have diverse origin. In [18], it is mentioned that the concatenation of features did not help. They also tried the combination of features using what they called the selection method and combination method, both in the segmentation and in the clustering phase. We mentioned in the baseline system how to combine MFCC features plus TDOA features combining them at the likelihood stage but without normalization as in [18] and both at the segmentation and clustering stage [20]. When using the ICC features, the $\mathbf{icc}$ vector is appended to the $\mathbf{tdoa}$ vector to form a joint second vector and follow the same combination strategy. When using the F0 features, these features make up a third stream with separate models for each cluster. The combination of all three streams is made in the same way as in (2) but now the combined likelihood for the $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ streams is

$$\begin{aligned}\log p(\mathbf{x},\mathbf{y},\mathbf{z}|\theta_a) = w_x \log p(\mathbf{x}|\theta_{ax}) \\ + w_y \log p(\mathbf{y}|\theta_{ay}) \\ + w_z \log p(\mathbf{z}|\theta_{az}) \qquad (4)\end{aligned}$$

[1] The conditions for the evaluation prevent us from specifying the authors and the rank of the systems presented, but they can be consulted in [26].

TABLE I
LIST OF MEETINGS FOR THE DEVELOPMENT SET USED IN THE EXPERIMENTS

		Meeting	# of microphones
1	devel06	AMI_20041210-1052	12
2		AMI_20050204-1206	16
3		CMU_20050228-1615	3
4		CMU_20050301-1415	3
5		ICSI_20000807-1000	6
6		ICSI_20010208-1430	6
7		LDC_20011116-1400	8
8		LDC_20011116-1500	8
9		NIST_20030623-1409	7
10		NIST_20030925-1517	7
11		VT_20050304-1300	2
12		VT_20050318-1430	2
13	RT06	CMU_20050912-0900	2
14		CMU_20050914-0900	2
15		EDI_20050216-1051	16
16		EDI_20050218-0900	16
17		NIST_20051024-0930	7
18		NIST_20051102-1323	7
19		VT_20050623-1400	4
20		VT_20051027-1400	4
21	RT07	CMU_20061115-1030	3
22		CMU_20061115-1530	3
23		EDI_20061113-1500	16
24		EDI_20061114-1500	16
25		NIST_20051104-1515	7
26		NIST_20060216-1347	7
27		VT_20050408-1500	4
28		VT_20050425-1000	7

in which similarly to (2) θ_{az} is the model created for any given cluster a using the stream $\mathbf{z}$. We have also tried the same strategy but using four streams $\mathbf{mfcc}$, $\mathbf{tdoa}$, $\mathbf{icc}$ and $\mathbf{lif0}$.

Fig. 2 shows the architecture of the proposed system. Blocks in the dashed line in the picture represent the modules of the proposed new system.

IV. Corpora and Evaluation Measures

In this paper, a subset of the NIST Rich Transcription of 2002–2005 sets, the RT06 set and the RT-07 set has been used as the development set. For the evaluation set we have used RT-90 set [15]. A subset of 12 meetings from RT02, RT04 and RT05 (called devel06 in [20]) together with RT06 and RT07 (called—DEVELSET—from now on) is made up of more than eighteen hours of audio data divided into twenty eight different meetings (see Table I), and RT-09 comprises more than five hours of audio data divided into seven different meetings.

The segments (UEM parts) defined by NIST for the official evaluations have been used to measure the performance of the systems described in this work. These parts consist of 27 612.64 seconds (2 761 264 frames) for the—DEVELSET—set and 10 858.49 seconds (1 085 849 frames) for RT-09 set that are taken into account to calculate the statistical significance of the results.

The speaker diarization performance is evaluated by comparing the hypothesis segmentation, given by the system, with the reference segmentation provided by NIST [15]. This reference segmentation was generated by hand according to a set of rules also defined by NIST and the exact speaker change points are calculated by force aligning the head mounted microphone audio to the reference transcripts using tools facilitated by the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI). In the evaluation plan the evaluation metric and a program to calculate it from both transcriptions is also defined. The error obtained is called the diarization error rate (DER) and it takes three errors into account (miss, false alarm, and speaker error). The error is time-based. A miss error occurs when a speech segment is classified as non-speech or an overlapping speaker is missing in the hypothesis. A false alarm (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.

V. Experiments

A. Preliminary Experimentation With ICC Features

Experiments and discussion of results with ICC features have been presented in another paper [21], so in this paper we will only give a summary for completeness. For these experiments the all06 set has been used.

In Fig. 4, DER is shown versus the weight applied to the MFCC stream for the all06 (devel06 + RT06) set for three systems, the baseline system ($\mathbf{mfcc}$ in a first stream and $\mathbf{tdoa}$ in a second stream) the proposed improvement $\mathbf{mfcc}$ plus $\mathbf{tdoa} + \mathbf{icc}$ in a second stream and an alternative system using three streams, $\mathbf{mfcc}$, $\mathbf{tdoa}$ and $\mathbf{icc}$ for which the $\mathbf{tdoa}$ vector and the $\mathbf{icc}$ vector are given the same weight. The baseline system has a DER for this set of 13.4% which has been outperformed by joining TDOA and ICC in the same vector thus obtaining a DER of 12.7%. A significant 5.2% relative reduction in DER was obtained. This experiment demonstrated that ICC features can be successfully incorporated in an improved speaker diarization system. The alternative system obtains an error of 12.97% and also improves the baseline but not as much as the proposed system. Further research is needed to determine the reasons of this behavior, one of them being that there is a strong correlation between TDOA features and ICC features, both of them related to the location of the speaker. As we will see later in Section V-C when using four sets of features, best results are obtained by concatenating TDOA features and ICC features instead of using them separately.

Experiments with the DEVELSET including the ICC features render a 4.6% relative DER improvement (13.04% versus 13.67%) over the baseline system that does not use ICC features (see Table II lines 2 and 4). For the RT09 set, the ICC features render a significant relative DER improvement of 7.9%.

B. Preliminary Experimentation With F0 Features

We made preliminary experimentation using plain interpolated F0 features $\mathbf{if0}$ and the logarithm of the interpolated F0

TABLE II
DER FOR EXPERIMENTS FOR THE BASELINE, AND FOR THE EXPERIMENTS INCLUDING TDOA + ICC FEATURES AND FOR EXPERIMENTS INCLUDING TDOA + ICC FEATURES AND F0 FEATURES IN THREE STREAMS

System	DEVELSET (28 meetings)	Relative DER improvement from the baseline	RT 09 (7 meetings)	Relative DER improvement from the baseline
Baseline	13.67 % ±0.05%		25.67%±0.11%	
Baseline plus **lif0** (3 streams)	12.67% ±0.05%	7.31%	22.94%±0.11%	10.63%
Baseline plus **icc** (2 streams)	13.04% ±0.05%	4.6%	23.64%±0.11%	7.9%
Baseline plus **icc** plus **lif0** (**UPM RT09 official system**)	**11.57%±0.05%**	**15.36%**	**21.38%±0.10%**	**16.71%**

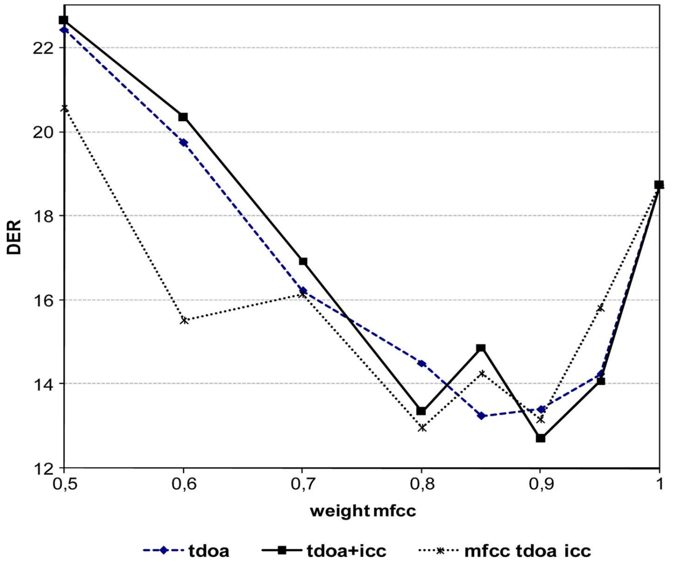

Fig. 4. DER for the all06 meetings set as a function of the weight associated to the first stream used by the system (always the stream with the MFCC features). The dashed line establishes the DER baseline to be improved (DER obtained using $\mathbf{mfcc}$ and $\mathbf{tdoa}$ streams). The $\mathbf{tdoa}+\mathbf{icc}$ graph uses as a second stream the concatenation of the $\mathbf{toda}$ and $\mathbf{icc}$ vectors. The $\mathbf{mfcc}$ $\mathbf{toda}$ $\mathbf{icc}$ graph uses three streams for which the weights for $\mathbf{toda}$ and $\mathbf{icc}$ streams are the same.

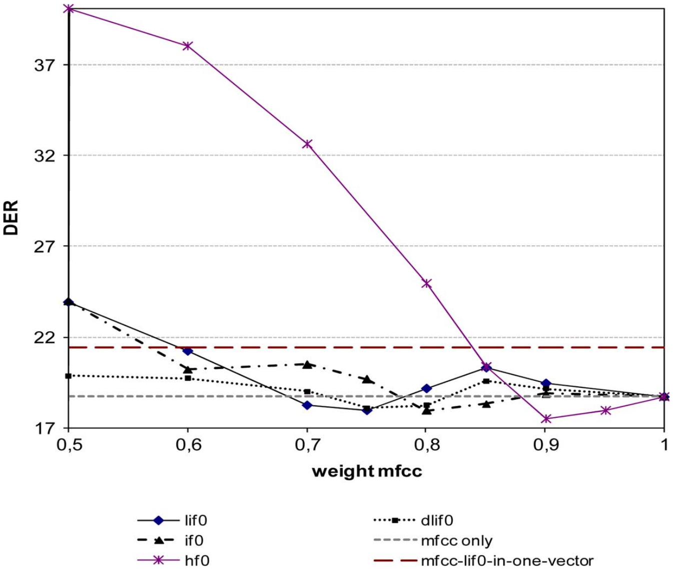

Fig. 5. DER for a system that mixes $\mathbf{mfcc}$ with either $\mathbf{lif0}$, $\mathbf{if0}$, $\mathbf{dlif0}$ or $\mathbf{hf0}$ as a function of the weight applied to the $\mathbf{mfcc}$ vector for the all06 set. In the picture DER for $\mathbf{mfcc}$ only features and $\mathbf{mfcc}$ and $\mathbf{lif0}$ features concatenated in the same vector are also presented.

features $\mathbf{lif0}$, the differential logarithm F0 features $\mathbf{dlif0}$ and the histogram F0 features $\mathbf{hf0}$ starting with 1 GMM (Gaussian mixture model) per cluster. These features have been combined with MFCC features to create an experimental system $\mathbf{mfcc}$ plus either one of the other features. The DERs for the all06 set across the weights for the $\mathbf{mfcc}$ stream using either $\mathbf{lif0}$ $\mathbf{if0}$, $\mathbf{dlif0}$ or $\mathbf{hf0}$ are presented in Fig. 5.

In Fig. 5, it can be seen that there are several weighting points in which the F0 features improve the MFCC features, thus confirming that the F0 adds information to the MFCC features. The absolute minimum is obtained by using the F0 histogram $\mathbf{hf0}$ but for the neighboring weighting points the DER increases quite abruptly. The next minimums are obtained using either the interpolated logarithm of F0 or the plain interpolated F0. The question is whether this F0 information in any of its forms can be used in an MDM system which also combines information from localization features. This will be shown in the experiments in the next section. In Fig. 5, the DER obtained is also represented when the MFCC features are concatenated with the $\mathbf{lif0}$ features in a single vector. The results degrade, thus confirming the experiments in [18]. The nature of F0 is quite different from the MFCC coefficients and the concatenation of both features does not help. The fact that we use a diagonal covariance matrix may have an influence on this result.

C. Experiments With the Combination of All Features

We made experiments using the DEVELSET combining different F0 methods and the baseline system. The results can be consulted in Table III.

It can be seen that the use of the lif0 stream delivers the best performance. The absolute minimum obtained in the previous experiment using hf0 has not been maintained in the new experiments, possibly due to the fact that the minimum obtained with these features is very unstable because the neighbors of the minimum in Fig. 5 have a much greater DER.

Finally, in Table II we present the results of the baseline system ($\mathbf{mfcc}$ plus $\mathbf{tdoa}$, second row), and the improvements obtained from the baseline by including the $\mathbf{lif0}$ stream (third row). By including the $\mathbf{lif0}$ we have been able to improve the

TABLE III
DER FOR THE DEVELSET USING DIFFERENT WAYS OF CALCULATING F0. THE BASELINE IS THE BEST PREVIOUS SYSTEM THAT COMBINES TDOA AND ICC FEATURES

System	DEVELSET
Baseline plus **icc**	13.04%±0.05%
Baseline plus **icc** plus **lif0**	**11.57%±0.05%**
Baseline plus **icc** plus **if0**	13.39%±0.05%
Baseline plus **icc** plus **dlif0**	14.01%±0.05%
Baseline plus **icc** plus **hf0**	16.98%±0.06%

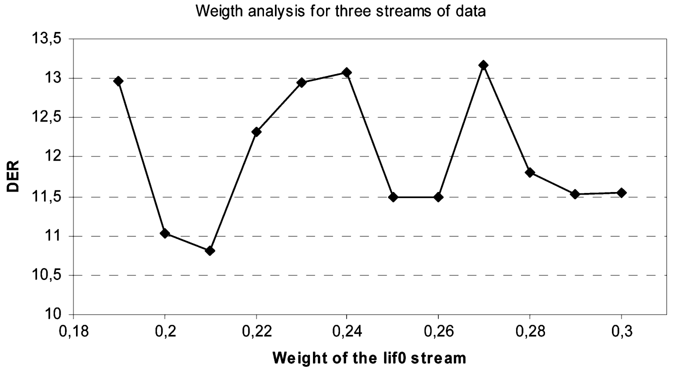

Fig. 6. DER for the all06 set for the three stream system versus different weight values for the **lif0** stream. The remaining weight (up to 1) is divided between the other two streams keeping a ratio between **mfcc** weights and **tdoa** + **icc** weights of 9.

TABLE IV
DER RESULTS INCLUDING ICC AND F0 FEATURES USING FOUR SEPARATE STREAMS

System	DEVELSET (28 meetings)	RT 09 (7 meetings)
Baseline	13.67 % ±0.05%	25.67%±0.11%
Baseline plus **icc** plus **lif0- (UPM- RT09 contrastive system)**	**12.12%±0.05%**	**22.43%±0.11%**
Relative improvement	11.33%	12.62%

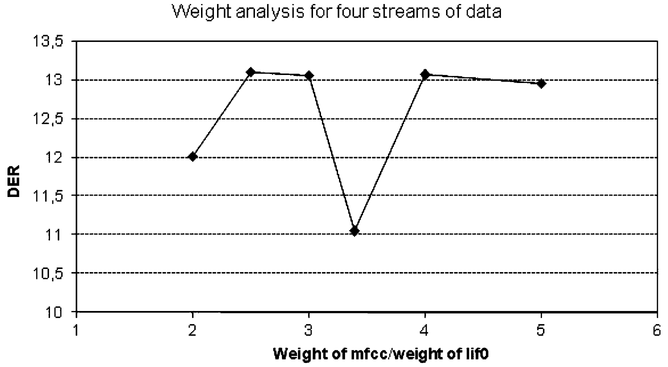

Fig. 7. DER for the all06 set for the four stream system versus different weight values for the quotient **mfcc**/**lif0** stream. The remaining weight (up to 1) is divided between the other two streams keeping a ratio between **mfcc** weight and **tdoa** weight of 9 and **mfcc** weight and **icc** weight of 9.

DER of the baseline by 7.31% and 10.63% relative for the DEVELSET and RT09 set, respectively.

We also present the improved baseline system **mfcc** plus **tdoa** + **icc** and the improvements obtained by including the **lif0** stream as a third stream. From this new baseline the relative improvements obtained by including F0 are 11.27% and 9.56% for the DEVELSET and RT09 set, respectively.

In Fig. 6, we show DER for the all06 set for the three streams system (best system) across different weights for the **lif0** stream. For the other two streams we use the strategy to divide the remaining weight (up to 1) between them keeping a weight ratio between **mfcc** and **tdoa** + **icc** of 9 (as it has been concluded from the experiments in Fig. 4). The minimum has been obtained for a **lif0** stream weight of 0.21 that corresponds to 0.711 and 0.079 weights for **mfcc** and **tdoa** + **icc**, respectively.

The final system with all the improvements together was presented at the RT09 evaluation in April 2009 obtaining a 21.38% DER on the RT 09 set. The relative improvements from the baseline are 15.36% and 16.71% for the DEVELSET and the RT09 set, respectively.

We also tried using four different streams, separating the TDOA features and the ICC features into two different vectors. The experiments using the same set of features but in separate streams, i.e., mfcc, tdoa, icc, and lif0 are shown in Table IV. The weights used in this case are 0.659, 0.073, 0.073, and 0.194 for mfcc, tdoa, icc, and lif0, respectively, which corresponds to a ratio of $\mathrm{weight}_{\mathrm{mfcc}}/\mathrm{weight}_{\mathrm{tdoa}} = \mathrm{weight}_{\mathrm{mfcc}}/\mathrm{weight}_{\mathrm{icc}} = 9$, and $\mathrm{weight}_{\mathrm{mfcc}}/\mathrm{weight}_{\mathrm{lif0}} =$ the same that the optimum that was obtained for the three streams case $= 0.339$. Other values for $\mathrm{weight}_{\mathrm{mfcc}}/\mathrm{weight}_{\mathrm{lif0}}$ keeping the other two ratios constant $= 9$ were tested with a subset of the database obtaining lower performance (see Fig. 7). This system was presented as a contrastive system in RT09 evaluation obtaining 22.43% DER. It can be seen that although they improve also the results of the baseline, the improvement is not as big as the improvement in the previous (official) system (see the discussion section for an explanation).

VI. DISCUSSION

We have proved that both ICC features and F0 features improve system performance. The ICC features do improve the baseline system.

Using the baseline and using the baseline plus the ICC features it is demonstrated that the F0 features can be combined with other features for speaker diarization. Instead of testing F0 with SDM meetings as in previous experiments [18] we have successfully integrated it into an MDM system including both ICC features and F0 features and obtained a significant relative improvement of 15.36% and 16.71% for the development set and the evaluation set, respectively. Since the features are quite independent of other modules of the system, we think that these new features could be incorporated into other state of the art systems.

In Table V, we present overall results for RT-09 meeting by meeting. By comparing the third and sixth columns it can be seen that with all of the contributions included there are significant improvements for six of the meetings and no changes for one of them (which is also the meeting with lowest DER and lowest speaker error). In the last column the speech/non-speech errors are presented for all the meetings and systems. It can

TABLE V
DETAILED %DER RESULTS COMPARING BASELINE SYSTEMS AND THE IMPROVEMENTS FOR RT09. THE LAST COLUMN SHOWS THE SPEECH/NON-SPEECH ERRORS FOR ALL OF THEM

MEETING	# mic.	**baseline**	**baseline + lif0**	**baseline + icc**	**baseline + icc + lif0**	SPNSP ERROR (all systems)
EDI 20071128-1000	24	7.79	7.81	7.70	7.71	7.3
EDI 20071128-1500	24	55.85	33.89	55.85	33.91	12.2
IDI 20090128-1600	8	11.39	11.01	11.39	11.01	4.8
IDI 20090129-1000	8	18.6	15.38	18.6	15.38	9.6
NIST 20080201-1405	7	61.85	54.25	61.02	55.06	19.3
NIST 20080227-1501	7	11.87	18.00	11.94	11.38	8.8
NIST 20080307-0955	7	32.83	38.65	19.45	32.21	4.8
All		25.67±0.05	22.94±0.11	23.64±0.05	21.38±0.10	12.6
DER improvement over the baseline			10.63	7,9	16.71	
Speaker error improvement			20.88	15.51	32.83	

TABLE VI
NUMBER OF IDENTIFIED SPEAKERS, MISS SPEAKERS, AND FALSE ALARM SPEAKERS FOR RT09 AND ALL THE SYSTEMS TESTED

MEETING	**baseline**			**baseline + icc**			**baseline + lif0**			**baseline + icc + lif0**		
	ID_SPK	MISS	FA	ID_SPK	MISS	FA	ID_SPK	MISS	FA	ID_SPK	MISS	FA
EDI_20071128-1000	4		1	4		1	4		1	4		1
EDI_20071128-1500	3	1		3	1		4			3	1	1
IDI_20090128-1600	4		2	4		1	4		2	4		2
IDI_20090129-1000	4			4			4			4		
NIST_20080201-1405	4	1		4	1		5			5		
NIST_20080227-1501	6			6			6			6		
NIST_20080307-0955	5	6		6	5		5	6		6	5	
ALL	30	8	3	31	7	2	32	6	3	32	6	4

be noticed that a big part of the remaining errors are due to the speech/non-speech errors (both Miss and False alarms). If we do not take those errors into account the proposed system with the new features improves the speaker error (SPKR) by 20.88% and 32.83% relative for the DEVELSET and the RT09 set, respectively.

In Table VI, a detailed analysis of the number of identified speakers (ID_SPK), missed speakers (MISS), and false alarm speakers (FA) is presented meeting per meeting. It can be seen that compared to the original baseline, in the proposed final system the number of identified speakers augments by two (30 to 32) while the number of miss speakers decreases by two (8 to 6) although one false alarm speaker is added (3 to 4).

It is not easy to determine the method to mix both features, ICC and F0 to improve a system, *a priori*, since ICC features are related to the localization of the speakers thus becoming more independent of MFCC and F0 but not as much from TDOA, the joint modeling of ICC and TDOA makes more sense than modeling them separately. A canonical correlation analysis between TDOA features and ICC features for all the meetings in the all06 set renders an average value of 0.37, which is significant. A similar average correlation value of 0.35 between MFCC and F0 was obtained that would justify the joint modeling of these two features. However, this was not supported by the experimental results as can be seen from Fig. 5 (corroborating other published experiments [18]).

Experiments using four streams instead of three streams resulted in a lower relative improvement. However, no exhaustive search has been done with the four streams system mainly due to computation costs.

The number of initial Gaussians used in the model may have also some influence. We have used five Gaussians for the MFCC features and one Gaussian for the other features but thorough investigation on it has not been done. Further research will be needed to create algorithms that automatically determine the best way to combine all the features. For instance in [28], the authors combine MFCC features and TDOA features using an information theoretic combination that is based in a different diarization methodology [3].

VII. CONCLUSION

In this paper, we present the contributions from the UPM to the RT09 evaluation. We have proposed a new energy-related feature, named ICC which represents an improvement of a previously used localization vector (the TDOA vector). We also present an innovative method to use F0 successfully for the first time at the clustering stage of MDM meetings. Instead of normalizing the features across clusters and using only them in the segmentation phase [18], or using a long-term window [16], we have used a short term window and have applied them both to the segmentation and to the clustering stage obtaining improved results from two different baseline systems. The accumulated relative improvements using both ICC and F0 rise up to 15.36% and 16.71% for the development and testing set, respectively. If we consider only the speaker errors, the improvements of the proposed features are of 23.4% and 32.83% relative for the mentioned sets.

Acknowledgment

The authors would like to thank ICSI in providing the baseline system. They would also like to thank J. Ferreiros for technical discussions and other members of the Speech Technology Group and the help of Juan M. Montero with managing computer resources. They would also like to thank M. Hallet for the English revision of this paper.

References

[1] S. Tranter and D. Reynolds, “An overview of automatic speaker diarization systems,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[2] J.-H. Chang, N. S. Kim, and S. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[3] D. Vijayasenan, F. Valente, and H. Bourlard, “An information theoretic approach to speaker diarization of meeting data,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1382–1393, Sep. 2009.

[4] E. El-Khoury, C. Senac, and J. Pinquier, “Improved speaker diarization system for meetings,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP’09*, Apr. 2009, pp. 4097–4100.

[5] C. Wooters and M. Huijbregts, “The ICSI rt07s speaker diarization system,” *Lecture Notes in Computer Science*, vol. 4625, pp. 509–519, 2008.

[6] H. Sun, T. L. Nwe, B. Ma, and H. Li, “Speaker diarization for meeting room audio,” in *Proc. Interspeech*, Sep. 2009, pp. 900–903.

[7] C. Fredouille, S. Bozonnet, and N. Evans, “The LIA-Eurecom rt 09 speaker diarization system,” in *Proc. Rich Transcript. 2009 Meeting Recognition Eval. Workshop*, 2009.

[8] S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in *Proc. DARPA Speech Recognition Workshop*, 1998.

[9] T. H. Nguyen, E.-S. Chng, and H. Li, “T-test distance and clustering criterion for speaker diarization,” in *Proc. Interspeech*, Sep. 2008.

[10] C. Barras, X. Zhu, S. Meignier, and J. L. Gauvain, “Improving speaker diarization,” in *Proc. DARPA RT04*, Palisades, NY, Nov. 2004.

[11] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[12] D. P. W. Ellis and J. C. Liu, “Speaker turn segmentation based on between-channel differences,” in *Proc. NIST Meeting Recognition Workshop at ICASSP’04*, 2004.

[13] J. M. Pardo, X. Anguera, and C. Wooters, “Speaker diarization for multimicrophone meetings using only between-channel differences,” *Lecture Notes in Computer Science*, vol. 4299/2006, pp. 257–264, 2006.

[14] J. M. Pardo, X. Anguera, and C. Wooters, “Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences,” in *Proc. ICSLP*, Sep. 2006, pp. 2194–2197.

[15] “Rich transcription evaluation project,” National Institute of Technology (NIST), 2002–2009 [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt

[16] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, “Prosodic and other long-term features for speaker diarization,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 5, pp. 985–993, Jul. 2009.

[17] N. W. D. Evans, C. Fredouille, and J.-F. Bonastre, “Speaker diarization using unsupervised discriminant analysis of inter-channel delay features EURECOM,” in *Proc. ICASSP*, 2009, pp. 4061–4064.

[18] A. Gallardo-Antolin, X. Anguera, and C. Wooters, “Multi-stream speaker diarization systems for the meetings domain,” in *Proc. Interspeech*, 2006.

[19] G. Friedland, “·ICSI’s speaker diarization, submissions for RT’09,” presented at the NIST RT09 Evaluation, Jul. 24, 2010 [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/icsi_rt09.pdf

[20] J. M. Pardo, X. Anguera, and C. Wooters, “Speaker diarization for multiple-distant-microphone meetings using several sources of information,” *IEEE Trans. Computers*, vol. 56, no. 9, pp. 1212–1224, Sep. 2007.

[21] R. Barra-Chicote, J. M. Pardo, J. Ferreiros, and J. M. Montero, “Speaker diarization based on intensity channel contribution,” *IEEE Trans. Audio, Speech, Lang.*, vol. 19, no. 4, pp. 754–761, May 2011.

[22] M. S. Brandstein and H. F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” in *Proc. ICASSP*, 1997, pp. 375–378.

[23] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando, “Hybrid speech/non-speech detector applied to speaker diarization of meetings,” in *Proc. Speaker Odyssey*, 2006.

[24] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2003, pp. 411–416.

[25] X. Anguera, “Robust Speaker Diarization for Meetings,” Ph.D. dissertation, Univ. Politécnica de Catalunya, Barcelona, Spain, Oct. 2006.

[26] 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. 4299/2006, p. 319, 2006 [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2006/spring/pdfs/rt06s-SPKR-SAD-results-v5.pdf

[27] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995, ch. 14.

[28] D. Vijayasenan, F. Valente, and H. Bourlard, “An information theoretic combination of MFCC and TDOA features for speaker diarization,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 431–438, Feb. 2011.

José M. Pardo (M’84–SM’04) received the Telecommunication Engineering Degree (M.S.E.E.) and Ph.D. degrees from the Universidad Politécnica de Madrid, Madrid, Spain, in 1978 and 1981, respectively.

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 in 1983–1984, a Visiting Scientist at SRI International in 1986 and a Visiting Fellow at the International Computer Science Institute in 2005–2006. He has authored or coauthored more than 180 papers and holds two patents. His current research interests are speaker diarization, speaker modeling, pattern recognition and audio indexing.

Prof. Pardo won a National Award in 1980 for the best graduate in telecommunication engineering and a National Award for the Best Ph.D. Thesis in 1982. He was member of the ISCA Advisory Council from 1996 until 2006. He was chairman of Eurospeech 1995 and member of ELSNET Executive Board 1998–2004. He was member of NATO RSG 10 and IST 3 from 1994 to 2002. He is a member of ASA, ISCA, and EURASIP.

Roberto Barra-Chicote received the M.S.E.E. degree (with highest distinction) from the Technical University of Madrid, Madrid, Spain, in 2005.

In 2006, he was a Visitor Researcher with the Center for Spoken Language Research (CSLR), Colorado University. In 2008, he was a Visitor Researcher with the Centre for Speech Technology Research (CSTR), Edinburgh University. He has research interests in emotional speech synthesis, speaker diarization and emotional speech recognition, with around 30 refereed publications in these areas. He is a reviewer of several journals. He is an Assistant Professor at Universidad Politecnica de Madrid.

Rubén San-Segundo received the Ph.D. degree (with highest distinction) from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2002.

He did two summer stays in The Center of Spoken Language Research (CSLR), University of Colorado at Boulder, as visiting student. From September 2001 through February 2003, he worked at the Speech Technology Group of the Telefónica I + D. Currently, he is an Associate Professor in the Department of Electronic Engineering at UPM. He is the Coordinator of the Spanish Network on Speech Technologies (www.rthabla.es) and he has been vice-chair of the Special Interest Group of ISCA on Iberian Languages (http://www.il-sig.org/).

Ricardo Cordoba (M'00) received the Telecommunication Engineering Degree and Ph.D. degrees from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 1991 and 1995 respectively.

He has been a member of the Speech Technology Group, UPM, since 1990, teaching at UPM since 1993, working as Associate Professor since 2003. He has been an Associate Director of the Electronic Engineering Department, UPM, since 2008. He worked as Research Associate in the Speech, Vision, and Robotics Group, Cambridge University, U.K., in 2001. He has authored or coauthored more than 80 papers and holds one patent. His main topics of work are dialog systems and multimodality, language and speaker identification, and speech recognition and synthesis.

Beatriz Martínez-Gonzalez received the M.S.E.E. degree from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2010. She is currently pursuing the Ph.D. degree the Speech Technology Group, UPM under the supervision of J. M. Pardo.

Since 2009, she has been working in the Speech Technology Group, UPM. Her main research interest are speaker diarization and speaker modeling.