Speaker Diarization Based on Intensity Channel Contribution

Roberto Barra-Chicote, Jose Manuel Pardo, *Senior Member, IEEE*, Javier Ferreiros, *Senior Member, IEEE*, and Juan Manuel Montero, *Member, IEEE*

***Abstract*—The time delay of arrival (TDOA) between multiple microphones has been used since 2006 as a source of information (localization) to complement the spectral features for speaker diarization. In this paper, we propose a new localization feature, the intensity channel contribution (ICC) based on the relative energy of the signal arriving at each channel compared to the sum of the energy of all the channels. We have demonstrated that by joining the ICC features and the TDOA features, the robustness of the localization features is improved and that the diarization error rate (DER) of the complete system (using localization and spectral features) has been reduced. By using this new localization feature, we have been able to achieve a 5.2% DER relative improvement in our development data, a 3.6% DER relative improvement in the RT07 evaluation data and a 7.9% DER relative improvement in the last year's RT09 evaluation data.**

***Index Terms*—Intensity channel contribution (ICC), 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.

In some speech research areas, like automatic language identification, automatic speaker identification and verification or text-to-speech synthesis, state-of-the-art algorithms are evaluated over a common framework, in order to learn the goodness and weakness of each algorithm in comparison with the others. These evaluations [1]–[3], have contributed to a rapid improvement in the technology in those areas. In case of speaker diarization, NIST evaluations for meetings started in 2002 and have been held after that in 2004, 2005, 2006, 2007, and 2009 [4].

In [5], a detailed overview of automatic speaker diarization systems is given. 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. The accuracy of the VAD module is essential for the purity of the speech frames that would be used by the clustering module. VAD algorithms differ, depending on the type of non-speech sounds that appear next to the speech or mixed with it, from Gaussian mixture models (GMMs) to Laplacian and gamma probability density functions [6]. If speech is mixed with music, some authors have used extra features as the modulation energy [7]. Other authors have detected speech by using integrated algorithms that treat VAD and speaker diarization simultaneously [8]. Some speaker diarization systems use bottom-up agglomerative clustering [9], [10], while others use a top-down universal background model (UBM) as a starting point to apply iteratively adaptation techniques to build the speaker models [11]. Clustering algorithms are typically based on the Bayesian information criterion (BIC) distance [12] although recent studies have also presented great improvements using other alternatives based on the t-test distance [13]. Most systems extract spectral features related to the spectral envelope such as the Mel frequency cepstral coefficients (MFCCs) [9], [14], although some studies have presented improvements with the fusion of spectral envelope and pitch features [15]. In [16], an exhaustive analysis of the goodness of prosodic and long-term features in speaker diarization is presented.

In speaker diarization with multiple distant microphones (MDMs), redundant information is available (one signal per microphone) in comparison with single distant microphone (SDM) diarization. Commonly, all speech signals are combined into one [17], from which some acoustic features (such as spectral or prosodic features) are extracted. The other source of information used in MDM scenarios is the information related to speaker localization [18], such as the time delay of arrival (TDOA) features [19]. 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 [20], it was demonstrated that TDOA between channels could be combined with spectral features to obtain improved performance over a base system that only used spectral features. This TDOA information combined with the MFCC information has been used by all systems in the latest Rich Transcription evaluation [4].

The shortcomings of TDOA methods are because of distant microphones. There are noises and reverberation 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. In this paper, we propose a new source of information related

to speaker localization that complements current TDOA features and by extension the MFCC features. We propose to use the information on intensity channel contribution (ICC) to carry out speaker diarization by itself and combined with current systems. We demonstrate that by using this new feature, previous state-of-the-art systems can be improved.

The paper is organized as follows. Section II presents the database used and the experiments carried out, Section III presents the system architecture and the mechanisms for obtaining the new features, Section IV describes the system setup, Section V presents experiment results, Section VI is the discussion, and Section VII ends with the conclusion.

II. Corpora

In this paper, a subset of the NIST Rich Transcription of 2002–2005 and RT06 has been used for development and RT-07 has been used as the test set. Additionally, results are given for RT-09 [4].

A subset of RT02, RT04, and RT05 [named devel06 in [21] and RT-06 together (named all06 from now on)] is made up of more than 13 hours of audio data divided into 20 different meetings, and RT-07 comprises more than five hours of audio data divided into eight 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 16 793.56 seconds (1 679 356 frames) for all06 and 10 819 seconds (1 981 900 frames) for RT-07 and 10 858.49 seconds (1 085 849 frames) for RT-09 that are taken into account to calculate the statistical significance of the results.

III. System Description

A. System Architecture and Baseline Features

Fig. 1 shows the system architecture. The input coming from M different microphones ($\{x_m[k]\}$) 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 ($x_i[k]$, the one with highest SNR). Then the GCC-PHAT between $x_i[k]$ and $x_j[k]$ is estimated as

$$G_{PHAT}(f) = \frac{X_i(f)\,[X_j(f)]^*}{\left|X_i(f)\,[X_j(f)]^*\right|} \tag{1}$$

where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals and $[\,]^*$ denotes the complex conjugate. The TDOA for these two microphones is estimated as

$$TDOA = \hat{d}(i,j) = argmax_d\left(\hat{R}_{PHAT}(d)\right) \tag{2}$$

where $\hat{R}_{PHAT}(d)$ is the inverse Fourier transform of (1). The set of TDOAs from each microphone to the reference channel will form what we call the TDOA vector ($[tdoa]$).

Once the $[tdoa]$ vector is calculated, a weighted delay-and-sum algorithm is applied in the *Acoustic Fusion* module, where the input signals $\{x_m[cS+k]\}$ are delayed and added together with a triangular window $\alpha[k]$ to generate a new composed signal

$$\begin{aligned} y[cS+k] = \alpha[k] \cdot \sum_{m=1}^{M} w_m[c] \cdot x_m\left[cS + k - TDOA^m[c]\right] \\ + (1-\alpha[k]) \cdot \sum_{m=1}^{M} w_m[c] \\ \cdot x_m\left[cS + k - TDOA^m[c-1]\right] \end{aligned} \tag{3}$$

$y[cS+k]$ is the composed signal, c is the segment being processed, S is the segment length (we use 250 ms), k is the sample within the segment being processed, $TDOA^m[c]$ is the delay between channel m and the reference channel for segment c and $w_m[c]$ is a weight factor applied to the channel m of segment c that is dynamically calculated; see [17].

The composed signal is then processed by the *MFCC estimation* module, where MFCC vectors of 19 components ($[mfcc]$) are calculated every 30 ms using a window shift of 10 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 a 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 about the VAD module can be found in [23].

The segmentation and agglomerative clustering process consists of an initialization and a 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. Every sub-state is modeled using a Gaussian mixture model (GMM) containing a number of components that has to be specified initially. 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.

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 hypothesis to have the same number of parameters [24]. When all possible merge pairs give a negative Δ BIC the merging is stopped.

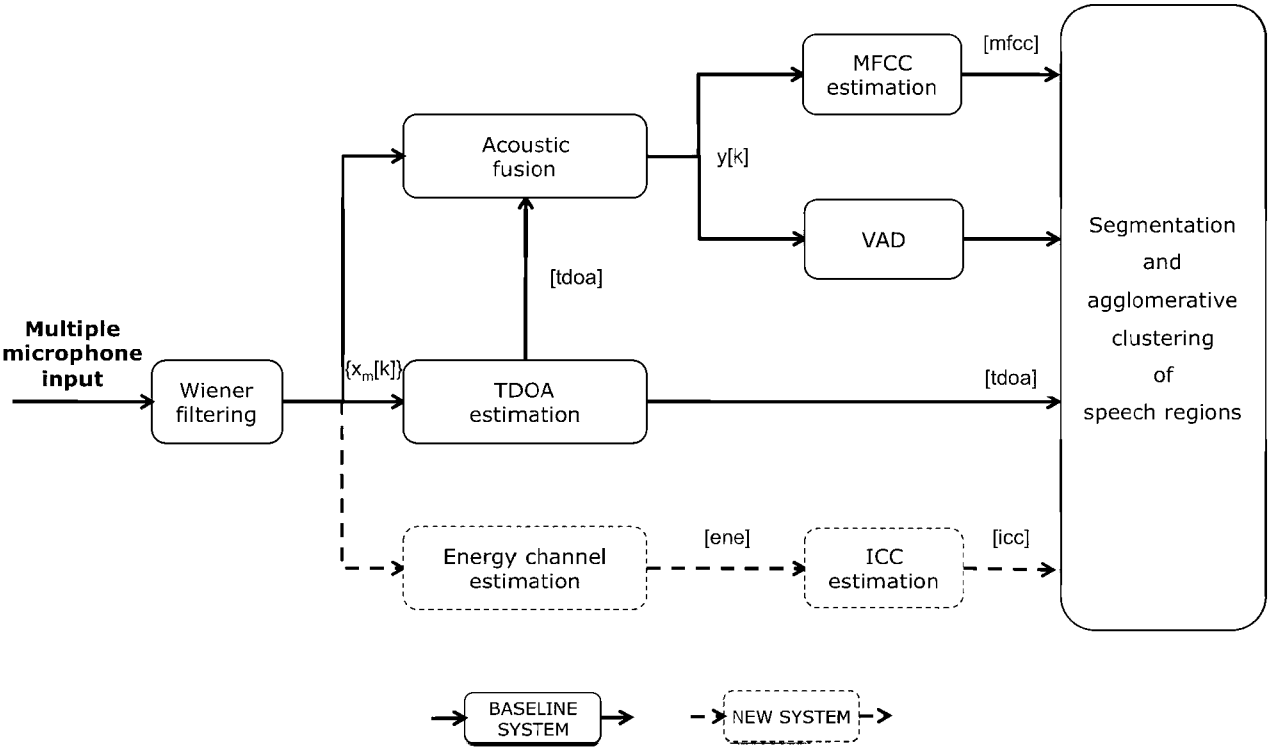

Fig. 1. Block diagram of the diarization system. The solid black flow diagram represents the baseline system, while the red/gray dash flow diagram represents the new modules and flow added to the system.

The baseline features used in the diarization task are the MFCC combined with the TDOA. In the implemented system [20], [21], the first 19 MFCC coefficients are extracted and treated as the $[x]$ stream and the TDOA features are treated as the $[y]$ stream. Each source of information is modeled using a statistical model whose individual likelihoods are combined using

$$logp([x],[y]|\theta_a) = w_x \cdot logp([x]|\theta_{ax}) + w_y \cdot logp([y]|\theta_{ay}) \quad (4)$$

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 $[x]$, and θ_{ay} is the model created for cluster a using the stream$[y]$.

B. Intensity Level Features

In this paper, we propose the use of intensity measures as an additional feature related to speaker localization. The red/gray dashed flow diagram in Fig. 1 shows the new modules added to the diarization system. The hypothesis proposed is based on the assumption that the position of the speaker has an impact on the speech intensity level captured by each microphone, and consequently to its location, analogous to the TDOA.

The energy captured by each channel is related to the distance of the speaker to that particular channel: when higher energy is detected, it means that the speaker is closer to that channel. 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 feature 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.

In order to confirm the hypothesis presented, two new features are considered:

- the absolute intensity $ene[n,m]$ of each audio channel m at frame n;
- the speech intensity channel contribution (ICC), $icc[n,m]$ the contribution of the absolute intensity per channel m at frame n to the sum of speech intensities coming from all the channels at frame n; see (5)

$$icc[n,m] = \frac{ene[n,m]}{\sum_{m=0}^{D_{\mathrm{ICC}}-1} ene[n,m]} \quad (5)$$

where D_{ICC} is the dimension of the ICC feature vector (equivalent to the number of microphones).

An initial oracle study was carried out using Multi-Dimensional Scaling (MDS) analysis over part of the devel06 set. MDS strategies [25] consider "proximity values" between two objects as their input (Euclidean distance between speakers model in our case). The result of the MDS analysis is a D_{MDS}-dimensional space in which each object is represented by a single point. MDS defines the localization of the points in the output space by minimizing the disparity between the Euclidean distances given the dissimilarity matrix (i.e., the proximity data) and the Euclidean distances in the object space, in the least squares sense" [26 p. 2169].

Using the speakers' references, we estimated a D_{TDOA}-dimensional Gaussian model (D_{TDOA} is the number of audio channels minus one) for each speaker using TDOA features and a D_{ICC}-dimensional Gaussian model using ICC features. An MDS analysis was conducted on speaker models to obtain a

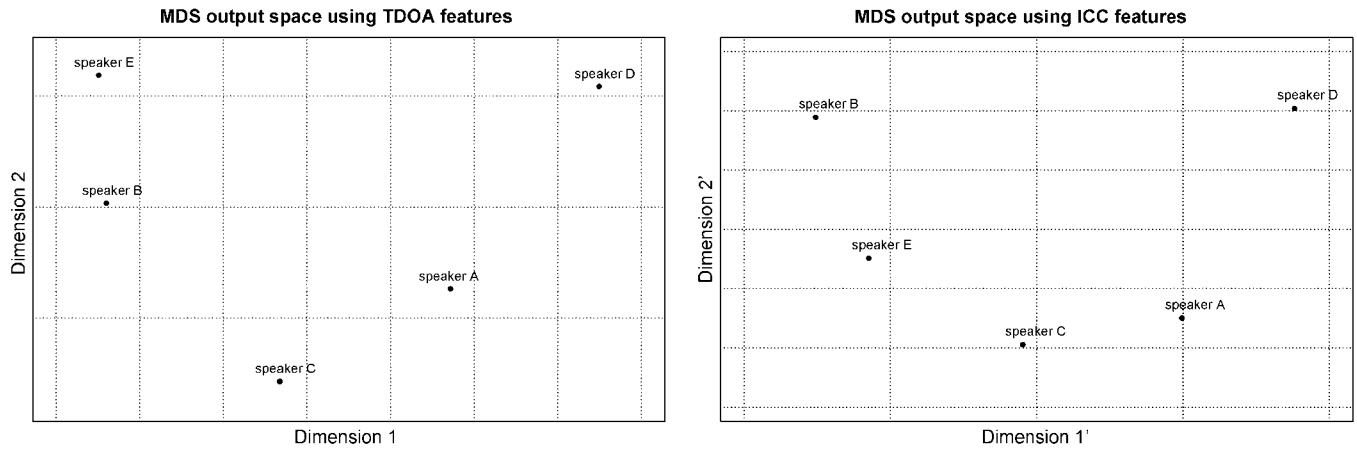

Fig. 2. Two-dimensional multi-dimensional scaling output space for the NIST20030623-1409 meeting using TDOA features and ICC features, respectively.

TABLE I
DER OBTAINED IN A SUBSET OF SIX RT-02 AND RT-04 MEETINGS, WHEN THE ABSOLUTE INTENSITY VALUE (ENE) FEATURES OF EACH AUDIO CHANNEL OR ICC FEATURES ARE USED

MEETING	ENE features $[ene]$	ICC features $[icc]$
ICSI_20000807-1000	59.66%	67.48%
ICSI_20010208-1430	59.25%	33.22%
LDC_20011116-1400	46.57%	34.55%
LDC_20011116-1500	51.53%	51.54%
NIST_20030623-1409	61.06%	6.32%
NIST_20030925-1517	40.66%	30.64%
ALL	$53.59\% \pm 0.16\%$	$37.40\% \pm 0.16\%$

TABLE II
DER OBTAINED IN THE ALL06 SET, WHEN USING MFCC, TDOA, ENE, OR ICC FEATURES SEPARATELY OR WHEN THE TDOA FEATURES ARE JOINED IN A COMMON STREAM WITH ENE OR ICC FEATURES

$[mfcc]$	$[tdoa]$	$[ene]$	$[icc]$
$18.73\% \pm 0.08\%$	$33.87\% \pm 0.09\%$	$55.65\% \pm 0.13$	$39.71\% \pm 0.10\%$

$[tdoa + ene]$	$[tdoa + icc]$
$35.36\% \pm 0.09$	$31.85\% \pm 0.09\%$

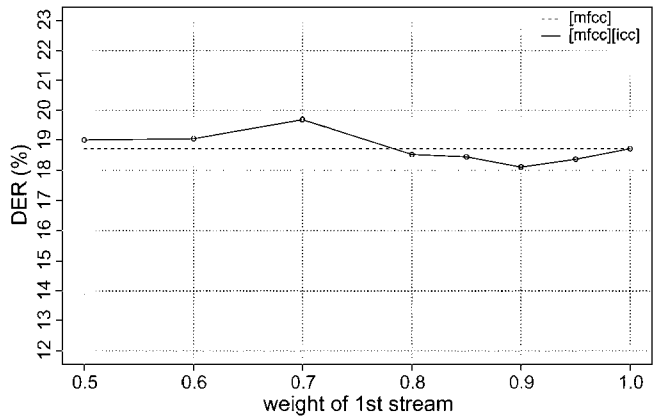

Fig. 3. DER as a function of the weight associated to the first stream used by the system (always the stream with the MFCC features). The dashed line establishes the DER baseline to be improved (DER obtained only using MFCC features). Results obtained on the all06 meetings set.

mapping of those models on a 2-D space with a goodness of fit measure (GoF) higher than 90%.

As an example, in Fig. 2 the 2-D MDS output space using TDOA and ICC features for the NIST_20030623-1409 meeting

TABLE III
DER OBTAINED IN THE ALL06 SET USING ONLY $[mfcc]$ FEATURES STREAM AND CONSIDERING $[icc]$ IN ADDITION TO $[mfcc]$. RESULTS OBTAINED WITH OPTIMAL WEIGHTS (0.9 AND 0.1) IN THE LAST CASE

$[mfcc]$ (baseline)	$[mfcc][icc]$	RDER (%)
$18.73\% \pm 0.08\%$	$18.12\% \pm 0.08\%$	3.3%

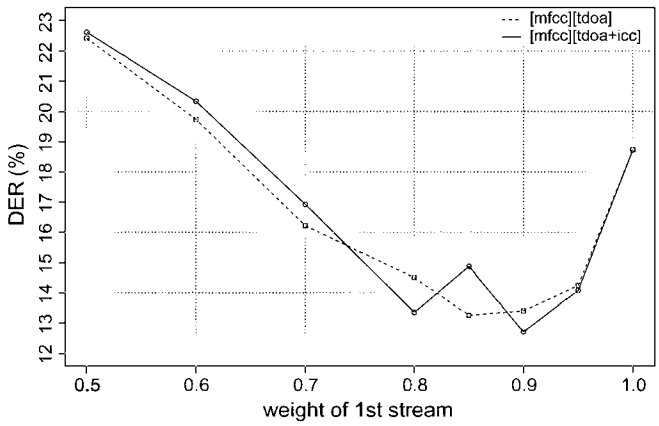

Fig. 4. DER as a function of the weight associated to the first stream used by the system (always the stream with the MFCC features). The dashed line establishes the DER baseline to be improved (DER obtained using $[mfcc]$ and $[tdoa]$ streams). Results obtained on the all06 meetings set.

TABLE IV
DER OBTAINED IN THE ALL06 MEETINGS SET USING $[mfcc]$ AND $[tdoa]$ FEATURES STREAMS AND CONSIDERING THE JOINED $[tdoa + icc]$ IN ADDITION TO $[mfcc]$. RESULTS OBTAINED WITH WEIGHTS (0.9 AND 0.1 IN BOTH CASES)

$[mfcc][tdoa]$ (baseline)	$[mfcc][tdoa + icc]$	RDER (%)
$13.40\% \pm 0.07\%$	$12.70\% \pm 0.07\%$	5.2%

can be seen. The MDS approach shows the distance between speakers, using TDOA or ICC features. One of the characteristics of the MDS analysis is the relationship (direct or indirect) of the output dimensions with some physical dimension related to the objects (coordinates in the room map in this case). The Pearson Coefficient between the speakers' localization in the two MDS is 0.86. This correlation is high, supporting a certain relationship between both features. Unfortunately, information on the geometry of the meeting room, information on the speakers and microphone localization, instant video in-

TABLE V
DER, USING THE OPTIMAL WEIGHTS, OBTAINED FOR EACH MEETING OF THE RT-07 MEETINGS SET

MEETING	$[mfcc]$	$[mfcc][tdoa]$	$[mfcc][icc]$	$[mfcc][tdoa+icc]$
CMU_20061115-1030	24.26%	22.15%	24.50%	24.02%
CMU_20061115-1530	11.85%	10.62%	12.00%	10.87%
EDI_20061113-1500	19.10%	20.42%	21.41%	13.06%
EDI_20061114-1500	21.74%	28.90%	20.89%	29.29%
NIST_20051104-1515	6.05%	5.39%	6.06%	5.36%
NIST_20060216-1347	11.42%	7.43%	12.16%	7.75%
VT_20050408-1500	7.53%	3.48%	7.58%	3.95%
VT_20050425-1000	18.49%	16.42%	20.66%	16.05%
AVERAGE DER	$14.91\% \pm 0.05\%$	$14.12\% \pm 0.05\%$	$15.49\% \pm 0.05\%$	$13.61\% \pm 0.05\%$

formation, etc. is not available for the evaluated corpora, not being possible to confirm empirically the relationship between the axes and the location of the speakers.

Further research with additional labeled corpora needs to be carried out to find an interpretation of the MDS output dimensions. However, in the last instance, the goodness of the proposed ICC features must be assessed using the diarization error rates (DER) which will be done in the next section.

IV. SYSTEM SETUP

A. Preliminary Experiments

In order to assess the intensity related features relevance, the DER has been calculated on a subset of six RT-02 and RT-04 meetings. Table I presents the results obtained when using the absolute intensity features ($[ene]$) and the ICC feature [as defined in (5)], respectively.

Results show that intensity level features carry information related to the location of the speaker. Much better results are obtained with ICC features (37.40%), that clearly outperform the results obtained with absolute intensity level features (53.59%).

Table II presents the results obtained on the all06 set when the system uses only one feature stream ($[mfcc]$, $[tdoa]$, $[icc]$, or $[ene]$), or when all information related to the speaker location ($[tdoa]$ and $[icc]$) is considered in the same stream (joint vector $[tdoa + icc]$ with dimension equal to $D_{\mathrm{TDOA}} + D_{\mathrm{ICC}}$). For completeness the results with the combination of $[tdoa]$ and $[ene]$ (joint vector $[tdoa + ene]$) are also presented. Separately, MFCC features clearly provide the best system performance (18.73%), surpassing TDOA features (33.87%) and ICC features (39.71%). However, when these last two features are jointly considered (31.85%) they outperform their separate performance indicating that ICC features can be combined with the TDOA information to improve performance. No improvement is obtained when the TDOA features are joined with absolute intensity features in the same stream (35.36%). This fact may well be explained by the fact that energy features alone provide bad results (55.65%) and instead of complementing the TDOA features performance, they degrade it. Note that although apparently both absolute energy and ICC features 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.

TABLE VI
DER, USING THE OPTIMAL WEIGHTS, OBTAINED FOR EACH MEETING OF THE RT-09 MEETINGS SET

MEETING	$[mfcc][tdoa]$	$[mfcc][tdoa+icc]$
EDI_20071128-1000	7.79%	7.70%
EDI_20071128-1500	55.85%	55.85%
IDI_20090128-1600	11.39%	11.39%
IDI_20090129-1000	18.60%	18.60%
NIST_20080201-1405	61.85%	61.02%
NIST_20080227-1501	11.87%	11.94%
NIST_20080307-0955	32.83%	19.45%
AVERAGE	$25.67\% \pm 0.11\%$	$23.64\% \pm 0.11\%$

B. Integrating ICC Features in the Final System

After the preliminary experiments, we have carried out research in a system that integrates all the three features, MFCC, TDOA, and ICC.

Our baseline system ($[mfcc][tdoa]$) is based on the MFCC and the TDOA features treated as separated streams. More information regarding the experimental framework using this configuration can be found in [21].

In this paper, the inclusion of the ICC features has been analyzed using two different alternatives:

- $[mfcc][icc]$: MFCC and ICC are considered as D_{MFCC} and D_{ICC}-dimensional separated streams, respectively;
- $[mfcc][tdoa+icc]$: TDOA and ICC are modeled jointly in one $D_{\mathrm{TDOA}} + D_{\mathrm{ICC}}$-dimensional stream.

It can be seen in Fig. 3 that there are several points in which the results obtained with MFCC alone can be improved by merging them with ICC features similar to what was presented in [21] with the merging of MFCC features with TDOA features. Table III presents the optimum results for weights 0.9 and 0.1, respectively. The relative reduction of the error rate compared to the use of MFCC alone is a significant 3.3%.

Finally in Fig. 4 the results of the all06 set are presented changing the weights for the first and second string using MFCC features as the first string and TDOA+ICC features as the second string. The baseline system has a performance of 13.4%[1] which has been outperformed by joining TDOA and ICC in the same vector. Table IV presents the optimum results. A significant 5.2% reduction in the DER has been obtained.

[1]Weights 0.9, 0.1, see [21].

TABLE VII
NUMBER OF IDENTIFIED SPEAKERS (ID SPK), FALSE ALARMS (FA), MISSES (MISS) FOR RT07 AND RT-09 DATA SETS. # CHAN MEANS NUMBER OF MICROPHONES, AND # SPK MEANS THE NUMBER OF ACTUAL SPEAKERS AT EACH MEETING

MEETING				$[mfcc][tdoa]$			$[mfcc][tdoa+icc]$		
CODE		# CHAN	# SPK	ID SPK	MISS	FA	ID SPK	MISS	FA
RT-07	CMU_20061115-1030	3	4	4	-	1	4	-	1
	CMU_20061115-1530	3	4	4	-	1	4	-	1
	EDI_20061113-1500	16	4	3	1	-	4	-	-
	EDI_20061114-1500	16	4	3	1	-	3	1	-
	NIST_20051104-1515	7	4	4	-	1	4	-	1
	NIST_20060216-1347	7	6	6	-	-	6	-	-
	VT_20050408-1500	4	5	5	-	-	5	-	-
	VT_20050425-1000	7	4	3	1	-	3	1	-
RT-09	EDI_20071128-1000	24	4	4	-	1	4	-	1
	EDI_20071128-1500	24	4	3	1	-	3	1	-
	IDI_20090128-1600	8	4	4	-	2	4	-	1
	IDI_20090129-1000	8	4	4	-	-	4	-	-
	NIST_20080201-1405	7	5	4	1	-	4	1	-
	NIST_20080227-1501	7	6	6	-	-	6	0	-
	NIST_20080307-0955	7	11	5	6	-	6	5	-
ALL		-	63	52	11	6	54	9	5

V. SPEAKER DIARIZATION OF RT-07 AND RT-09

To assess the validity of the previous analysis, we have calculated the DER for the evaluation set (RT07) using the optimum weights obtained in the previous experiments with the all06 set. It can be seen in Table V that the use of the energy together with the TDOA improves the previous system by a significant 3.6% relative.

This contribution has been part of the system presented by Universidad Politecnica de Madrid (UPM) at RT 2009. To further demonstrate the validity of this approach, recent post-eval experiments performed with the recently released RT09 set (used as a second evaluation set) gives a result of 23.64 DER with ICC features compared to 25.67% without the ICC features, a relative improvement of 7.9%. Results are shown in Table VI.

VI. DISCUSSION

From the presented experiments, it is demonstrated that the ICC features constitute an additional source of information that can be used in conjunction with the previously used TDOA features. Results show a consistent improvement in the development set and the two RT evaluation data sets (5.2%, 3.6%, and 7.9% relative on all06, RT-07 and RT-09, respectively). These improvements suggest that ICC features might contribute to the improvement of alternative state of the art systems evaluated in the 2009 Rich Transcription Evaluation.

While the ICC features alone do not outperform the TDOA features alone, the join vector ($[icc][tdoa]$) outperforms both of them. In addition, while the simple union of ICC with MFCC features ($[mfcc][icc]$) does not outperform the improvement obtained when TDOA features are joined with MFCC features ($[mfcc][tdoa]$) when both location features are combined with MFCC ($[mfcc][tdoa+icc]$) the overall DER decreases. This result proves that both location features complement themselves and are more robust than any of them separately.

If we analyze the results in detail meeting by meeting we notice that the combination of ICC features with TDOA features provide very large improvements in the EDI_20061113-1500 (RT-07) meeting (36% relative) and in the NIST_20080307-0955 (RT-09) meeting (41% relative) while the variation of DER in the other meetings is very small.

TABLE VIII
EDI_20061113-1500 AND NIST_20080307-0955 SPEAKER IDENTIFICATION RESULTS OBTAINED WITH $[mfcc][tdoa]$ AND $[mfcc][tdoa+icc]$ SYSTEMS, RESPECTIVELY

EDI_20061113-1500		
SPK REF	$[mfcc][tdoa]$	$[mfcc][tdoa+icc]$
fee097	MISSED	ID
fee100	ID	ID
mee098	ID	ID
mee099	ID	ID

NIST_20080307-0955		
SPK REF	$[mfcc][tdoa]$	$[mfcc][tdoa+icc]$
302	MISSED	ID
303	ID	ID
304	MISSED	MISSED
305	ID	ID
306	ID	ID
307	MISSED	MISSED
308	ID	ID
309	MISSED	MISSED
310	MISSED	MISSED
311	ID	ID
312	MISSED	MISSED

However, the big improvement in these two meetings is enough to result in an overall significant improvement.

What can be done is a detailed analysis of the errors that we have got. In Table VII, we show the number of microphones and the number of reference speakers for each meeting of RT-07 and RT-09 to search for relationships between these two variables. A Pearson Coefficient of 0.3 shows a low correlation between the number of channels and DER. Table VII also presents the number of identified speakers (ID SPK), missed speakers (MISS) and false alarms (FA) obtained with $[mfcc][tdoa]$ and $[mfcc][tdoa+icc]$. Results show that the inclusion of the ICC features reduces the FA (from 6 to 5) and the MISS (from 11 to 9) improving the number of identified speakers (from 52 to 54). These improvements come from the EDI_20061113-1500 meeting and the NIST_20080307-0955 meeting (the ones with higher relative improvement in DER).

We have also analyzed which speakers have been identified correctly and those missed with the $[mfcc][tdoa]$ system and the $[mfcc][tdoa+icc]$ system. The analysis is presented in Table VIII. Both systems present a similar behavior, so most

of the time both systems make the same errors. However, the $[mfcc][tdoa + icc]$ system identifies speaker $fee097$ and speaker 302 while $[mfcc][tdoa]$ is not able to identify these two speakers resulting in an overall improvement in these two meetings.

VII. Conclusion

We have researched additional localization features to improve the results of a previous speaker diarization system.

We have proposed the use of a new measure, the ICC, as an additional source of information. We have successfully merged this information with TDOA and MFCC features to obtain an enhanced system, better than our baseline system which only used MFCC and TDOA features.

The enhanced system improves the baseline system significantly by 5.2% relative in the development set, as well as 3.6% and 7.9% relative in the RT-07 and RT-09 evaluation sets, respectively, thus demonstrating in all cases the robustness of the approach.

Acknowledgment

The authors would like to thank other members of the Speech Technology Group for technical discussions and the help of R. San Segundo with managing computer resources.

References

[1] "Language Recognition Evaluation," National Institute of Technology (NIST), 2009 [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/lre/

[2] 'Speaker Recognition Evaluation," National Institute of Technology (NIST), 1997–2010 [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/

[3] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[4] Rich Transcription Evaluation Project National Institute of Technology (NIST), 2002–2009 [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt

[5] 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.

[6] 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.

[7] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP-97*, Apr. 1997, vol. 2, pp. 1331–1334.

[8] 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.

[9] C. Wooters and M. Huijbregts, "The ICSI rt07s speaker diarization system," in *Lecture Notes in Computer Sciences*, 2008, vol. 4625, pp. 509–519.

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

[11] C. Fredouille, S. Bozonnet, and N. Evans, "The LIA-EURECOM rt 09 speaker diarization system," in *Proc. Rich Transcription 2009 Meeting Recognition Evaluation Workshop*, 2009.

[12] 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.

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

[14] 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.

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

[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] 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.

[18] 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.

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

[20] 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.

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

[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 Recognition Understanding*, 2003, pp. 411–416.

[25] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, ser. Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills, CA: Sage, 1978.

[26] J. L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *J. Acoust. Soc. Amer.*, vol. 110, pp. 2167–2182, 2001.

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

Since 2003, he has been a member of the Speech Technology Group, Technical University of Madrid. In 2006, he was a Visitor Researcher of the Center for Spoken Language Research (CSLR) at Colorado University. In 2008, he was a Visitor Researcher of the Centre for Speech Technology Research (CSTR), Edinburgh University. His main research interests are related to emotional speech synthesis and automatic emotion identification.

Jose Manuel Pardo (M'84–SM'04) received the M.S.E.E. and Ph.D. degrees from the Universidad Politecnica de Madrid, Madrid, Spain, in 1978 and 1981, respectively.

He has been 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, Cambridge, in 1983-1984, a Visiting Scientist at SRI International in 1986, and a Visiting Fellow at ICSI-Berkeley in 2005–2006.

Prof. Pardo received the Best Graduate National Award in 1980 and a Best Ph.D. Thesis National Award in 1982. He was Chairman of Eurospeech 1995, member of the ISCA Advisory Council, ELSNET Executive Board, and NATO RSG 10 and IST 3.

Javier Ferreiros received the M.S.E.E. and Ph.D. degrees with highest distinctions from the Universidad Politecnica de Madrid (UPM), Madrid, Spain, in 1990 and 1996, respectively.

Since 1988, he has been a member of the Speech Technology Group at UPM, where he holds an Associate Professor position and currently is the Director for academic planning of the Escuela Técnica Superior de Ingenieros de Telecomunicación. From October 1999 to April 2000, he stayed at ICSI, Berkeley, CA, as a Visiting Researcher. His research interests focus on spoken dialog systems.

Juan Manuel Montero received the M.S.E.E. and Ph.D. degrees with highest distinctions from the Universidad Politecnica de Madrid (UPM), Madrid, Spain, in 1992 and 2003, respectively.

He spent seven months in the Speech Group, ICSI, Berkeley, CA. Currently, he is an Associate Professor in the Department of Electronic Engineering at UPM and has been a member of the Speech Technology Group since 1990.