Robust Speaker Identification for Meetings: UPC CLEAR'07 Meeting Room Evaluation System

Jordi Luque and Javier Hernando

Technical University of Catalonia (UPC)
Jordi Girona, 1-3 D5, 08034 Barcelona, Spain
luque@tsc.upc.edu

Abstract. In this paper, the authors describe the UPC speaker identification system submitted to the CLEAR'07 (Classification of Events, Activities and Relationships) evaluation. Firstly, the UPC single distant microphone identification system is described. Then the use of combined microphone inputs in two different approaches is also considered. The first approach combines signals from several microphones to obtain a single enhanced signal by means a delay and sum algorithm. The second one fuses the decision of several single distant microphone systems. In our experiments, the latter approach has provided the best results for this task.

1 Introduction

The CHIL (Computers in the Human Interaction Loop) project [1] has collected a speaker database in several smart room environments and has organized last two years the Evaluation Campaign to benchmark the identification performance of the different approaches presented. The Person IDentification (PID) task is becoming important due to the necessity of identify persons in a smart environment for surveillance or the customizing of services. In this paper the UPC acoustic person identification system and the obtained results in the CLEAR'07 evaluation campaign [2] are presented.

The CLEAR PID evaluation campaign has been designed to study the issues that cause important degradations in the real systems. One of them is the degradation of performance in terms of the amount of speaker data available for training and testing. In most of the real situations we do not have enough data to obtain an accurate estimation of the person model. Usually, the systems show a big drop in the correct identification rates from the 5 seconds to the 1 second testing conditions.

The second evaluation goal focus on the combination of redundant information from multiple input sources. By means of robust and multi-microphone techniques the different approaches deal with the channel and noise distortion because the far-field conditions. No a priori knowledge about the room environment is known and the multi-microphone recordings from the MarkIII

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 266–275, 2008.

array were provided to perform the acoustic identification, whereas, the previous evaluation only used one microphone in the testing stage. For further information about the Evaluation Plan and conditions see [3].

Two different approaches based on a mono-microphone technique will be described in this paper. The single channel algorithm is based on a short-term estimation of the speech spectrum using Frequency Filtering (FF), as described in [4], and Gaussian Mixture Models (GMM) [5]. We will refer it to as: Single Distant Microphone (SDM) approach. The two multi-microphone approaches try to take advantage of the space diversity of the speech signal in this task. The first approach makes use of a Delay and Sum (D&S) [6] algorithm with the purpose to obtain a enhanced version of the speech. The second approach profits the multi-channel diversity fusing three SDM classifiers at the decision level. The evaluation experiments show that the SDM implementation seems to be suitable to the task obtaining a good identification rate only outperformed by the decision-fusion (D&F) approach.

This paper is organized as follows. In section 2 the SDM baseline is described and the two multi-microphone approaches are presented. Section 3 describes the evaluation scenario and the experimental results. Finally, section 4 is devoted to provide conclusions.

2 Speaker Recognition System

Below we describe the main features of the UPC acoustic speaker identification system. The SDM baseline system and the two multi-channel approaches shared the same characteristics about the parameterization and statistical modelling, but they differ in the use of the multi-channel information.

2.1 Single Distant Microphone System

The SDM approach is based on a short-term estimation of the spectrum energy in several sub-bands. The scheme we present follow the classical procedure used to obtain the Mel-Frequency Cepstral Coefficients (MFCC), however in this approach instead of the using of the Discrete Cosine Transform, such as in the MFCC procedure [7], the log filter-bank energies are filtered by a linear and second order filter. This technique was called Frequency Filtering (FF) [4]. The filter we have used in this work have the transform frequency response:

$$H(z) = z - z^{-1} \tag{1}$$

and it's applied over the log of the filter-bank energies. By performing a combination of decorrelation and liftering, FF yields good recognition performance for both clean and noisy speech. Furthermore, this new linear transformation, unlike DCT, maintains the speech parameters in the frequency domain. This Filter is computationally simple, since for each band it only requires to subtract the log FBEs of the two adjacent bands. The first goal of frequency filtering is to decorrelate the output parameter vector of the filter

bank energies like cepstral coefficients do. Decorrelation is a desired property of spectral features since diagonal covariance matrices are currently assumed in this work [8].

A total of 30 FF coefficients have been used. In order to capture the temporal evolution of the parameters the first and second time derivatives of the features are appended to the basic static feature vector. The so called Δ and Δ-Δ coefficients [9] are also used in this work. Note that, for that filter, the magnitudes of the two endpoints of the filtered sequence actually are absolute energies [10], not differences. That are also employed to compute the model estimation as well as its velocity and acceleration parameters.

Next, for each speaker that the system has to recognize a model of the probability density function of the FF parameter vectors is estimated. These models are known as Gaussian Mixture Models (GMM) [5]. A weighed sum of size 64 was used in this work. Maximum likelihood model parameters were estimated by means of the iterative Expectation-Maximization (EM) algorithm. It is well known, the sensitive dependence of the number of EM-iterations in the conditions of few amount of training data. Hence, to avoid over-training of the models, 10 iterations were enough for parameter convergence in both training and testing conditions.

In the testing phase of the speaker identification system, firstly a set of parameters $\mathbf{O} = \{\mathbf{o}_i\}$ is computed from the testing speech signal. Next, the likelihood that each client model is calculated and the speaker showing the largest likelihood is chosen:

$$s = \arg\max_j \left\{ L\left(\mathbf{O}|\lambda_j\right) \right\} \tag{2}$$

where s is the score of the recognized speaker. Therefore, $L\left(\mathbf{O}|\lambda_j\right)$ is the likelihood that the vector $\mathbf{O}$ has generated by the speaker of the model λ_j.

2.2 Delay-and-Sum Acoustic Beamforming

The Delay-and-Sum beamforming technique [6] is a simple and efficient way to enhance an input signal when it has been recorded on more than one microphone. It does not assume any information about the position of the microphones or their placement.

If we assume the distance between the speech source and the microphones is enough far we can hypothesize that the speech wave arriving to each microphone is flat. Therefore, the difference between the input signals, only taking into account the wave path and without take care about channel distortion, is a delay of arrival due the different positions of the microphones with regard to the source. So if we estimate the delay between two microphones we could synchronize two different input signal in order to enhance the speaker information and reduce the additive white noise.

Hence given the signals captured by N microphones, $x_i[n]$ with $i = 0 \ldots N-1$ (where n indicates time steps) if we know their individual relative delays $d(0, i)$ (called Time Delay of Arrival, TDOA) with respect to a common reference

microphone x_0 , we can obtain the enhanced signal by adding together the aligned signals as follows:

$$y(n) = x_0[n] + \sum_{i=1}^{N-1} W_i x_i[n - d(0,i)] \tag{3}$$

The weighting factor W_i, which is applied to each microphone to compute the beamformed signal, was fixed to the inverse of the number of channels.

In order to estimate the TDOA between two segments from two microphones we have used the Generalized Cross Correlation with PHAse Transform (GCC-PHAT) method [11]. Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is defined as:

$$\hat{G}_{PHAT_{ij}}(f) = \frac{X_i(f)\left[X_j(f)\right]^*}{\left|X_i(f)\left[X_j(f)\right]^*\right|} \tag{4}$$

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

$$\hat{d}_{PHAT_{ij}}(d) = \arg\max_d \hat{R}_{PHAT_{ij}}(d) \tag{5}$$

where $\hat{R}_{PHAT_{ij}}(d)$ is the inverse Fourier transform of $\hat{G}_{PHAT_{ij}}(f)$, the Fourier transform of the estimated cross correlation phase. The maximum value of $\hat{R}_{PHAT_{ij}}(d)$ corresponds to the estimated TDOA. This estimation is obtained from different window size depending of the duration of the testing sentence (1s/5s/10s/20s). In the training stage, the same scheme is applied and we obtain the TDOA value from the training sets of 15 and 30 seconds. Note the difference in the window size in every TDOA estimation because the whole speech segment is employed. A total of 20 channels were used, selecting equispaced microphones from the MarkIII 64 array.

2.3 Multi-microphone Decision Fusion

In this approach we have implemented a multi-microphone system fusing three SDM classifiers, each of them as described in Section 2.1, working on three different microphone outputs. The microphones $4, 34$ and 60 from MarkIII array have been used. The SDM algorithms are applied independently to obtain an ID decision in matching conditions.

Although they shared the same identification algorithm, the three classifiers sometimes do not agree about the identification of the segment data because of the various incoming reverberation or other noises in the different microphones. In order to decide a sole ID from the classifier outputs, a fusion of decisions is applied based on the following easy voting rule:

$$\begin{array}{ll} \text{if} \quad ID_i \neq ID_j \quad \vee i,j \neq i & \text{select the central microphone ID} \\ \text{if} \quad ID_i = ID_j \quad \text{for some } i \neq j & \text{select } D_i \end{array} \tag{6}$$

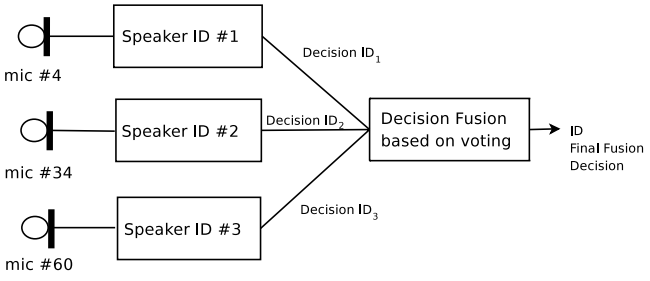

Fig. 1. Multi-microphone fusion, at the decision level, architecture

where ID_i is the decision of the classifier number i. In other words, an ID is decided if two of them agree, and the central microphone decision is chosen in the case all three classifier disagree. The selection of the central microphone decision is motivated by its better single SDM performance in our development experiments.

3 Experiments and Discussion

3.1 Database

A set of audiovisual far-field recordings of seminars and of highly-interactive small working-group seminars have been used. These recordings were collected by the CHIL consortium for the CLEAR 07 evaluation according to the "CHIL Room Setup" specification [1]. A complete description of the different recordings can be found in [3].

In order to evaluate how the duration of the training signals affects the performance of the system, two training conditions have been considered: 15 and 30 seconds, called train A and train B respectively. Test segments of different durations ($1, 5, 10$ and 20 seconds) have been used during the algorithm development and testing phases. There are 28 different personal identities in the database and a total of 108 experiments per speaker (of assorted durations) were evaluated.

For each seminar a 64 microphone channels, at 44.1 kHz and 16 bits/sample, were provided. Each audio signal was divided into segments which contain information of a sole speaker. These segments were merged to form the final testing segments (see the number of segments in Table 1) and the training sets A and B. The silences longer than one second were removed from the data. That is the reason why a speech activity detection (SAD) has been not used in the front-end of our implementations. The metric used to benchmark the quality of the algorithms is the percentage of correctly recognized people from the test segments.

Table 1. Number of segments for each test condition

	Number of segments	
Segment Duration	Development	Evaluation
1 sec	560	2240
5 sec	112	448
10 sec	56	224
20 sec	28	112
Total	756	3024

3.2 Experimental Set-Up

The database provided was decimated from 44.1KHz to 16KHz sampling rate. The audio was analyzed in frames of 30 milliseconds at intervals of 10 milliseconds. Each frame window was processed subtracting the mean amplitude and no preemphasis was applied to the signal. Next a Hamming window was applied to each frame and the FFT was computed. The corresponding FFT amplitudes were then averaged in 30 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. The microphone 4 from the MarkIII array was selected in the SDM algorithm with the purpose of comparing with the CLEAR'06 evaluation.

3.3 Results

In this section we summarize the results for the evaluation of the UPC acoustic system and the differences between the previous evaluation are examined. The Table 2 shows the correct identification rate obtained by the UPC acoustic implementations. That Table shows the rates obtained for the single microphone (SDM'07), Decision Fusion (D&F) and Beamforming (D&S) systems in either train A and train B conditions. Furthermore, the results from the single channel system from the previous evaluation (SDM'06) are also provided.

Some improvements have been performed on the system since the CLEAR'06 Evaluation, leading to better results than the ones presented in that. It can be seen that the results are better as the segments length increases. The Table 2 shows

Table 2. Percentage of correct identification obtained by the different UPC approaches

	Train A (15s)				Train B (30s)			
Duration	SDM'06	SDM'07	D&F	D&S	SDM'06	SDM'07	D&F	D&S
1s	75.04 %	78.6 %	**79.6** %	65.8 %	84.01 %	83.3 %	**85.6** %	72.2 %
5s	89.29 %	**92.9** %	92.2 %	85.7 %	97.08 %	95.3 %	**96.2** %	89.5 %
10s	89.27 %	**96.0** %	95.1 %	83.9 %	96.19 %	**98.7** %	97.8 %	87.5 %
20s	88.20 %	**98.2** %	97.3 %	91.1 %	97.19 %	**99.1** %	**99.1** %	92.9 %

this kind of behavior. In the SDM system, the results reach an improvement of up to 4.5% (absolute) in the recognition, comparing the train A with the train B condition.

On one hand, the decision fusion system seems, even with a very simple voting rule, to exploit the redundant information from each SDM system. This technique achieves the best results in the tests of 1s using any training set and in most of the test conditions of the training set B. On the other hand, as we can see in the Table 2, the Delay and Sum system does not provide good results to the task. The low performance of this implementation may be due to a not accurate estimation of the TDOA values. Other possible explanation could be the different background noise and the reverberation effects from the various room setups. The recordings was collected from 5 different sites, which could aid the GMM system to discriminate between the recorded speakers from the different room environments. As we can see in the Figure 2 mostly of the errors occurs between speakers of the same site.

In fact, neither of the systems presented in the evaluation based on any kind of signal beamforming did not show good results. By contrast, the same technique was applied in the Rich Transcription Evaluation'07 [12] obtaining good results in the diarization task.

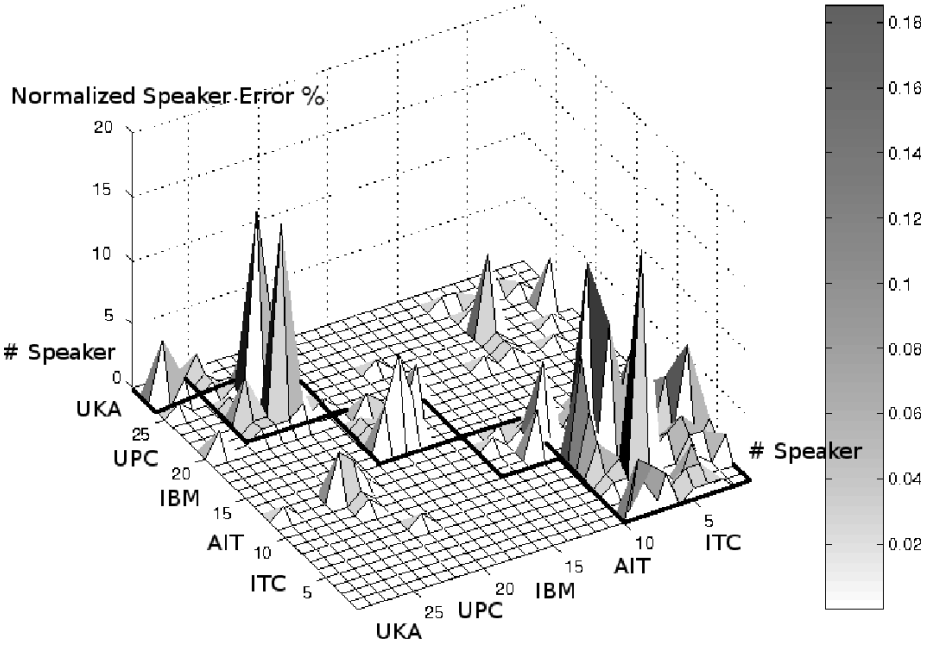

Fig. 2. Normalized Speaker Error from SDM in all test conditions. We can see the error mostly appears between the speakers of the same recording conditions.

The Figure 2 depicts the error behavior between speakers from the SDM implementation, a total of 348 over 3024 ID experiments. The boxes around the main diagonal enclose regions corresponding to speakers from the same site, that means, recordings with the same room conditions. As it has been commented above, we can see the number of speaker errors is higher around the main diagonal. The picture shows that the system mostly confuses the speakers from the same site. This kind of behavior could be due to the fact that our system is modelling both the speaker, accent or dialect, and the room characteristics, such as the space geometry or the response of the microphones.

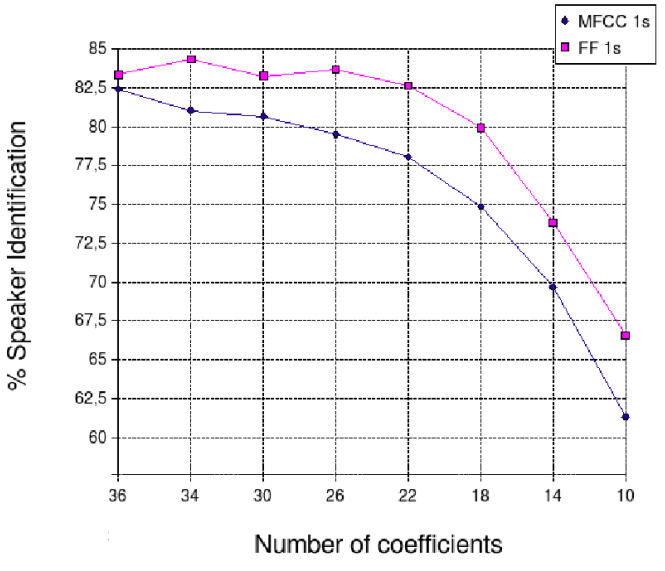

Fig. 3. Percentage of correct identification of the SDM approach in terms of the number of front-end parameters. The 30s training set and the 1s test condition from the Evaluation data 07 were employed to draw the figure.

3.4 Frequency Filtering Experiments

Some experiments were conducted focusing on the FF front-end. The Figure 3 shows the correct identification rate in terms of the number of parameters. Train A and 1s test condition have been selected to drawn the figure. Note that the number of coefficients are referred to the static parameters, but the total of parameters is three times more, including Δ and $\Delta - \Delta$. We can see that the optimum value of parameters, 34, is close to the value of 30 tuned during the development and applied in the submitted systems.In addition, the Figure 3 also shows the performance achieved by the MFCC coefficients, which always are lower than the FF results.

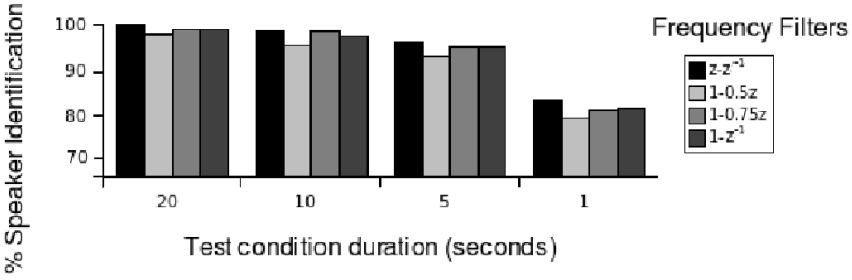

Fig. 4. Percentage of correct identification from the SDM approach using four different frequency filters.

Furthermore, a comparison between several frequency filters is provided in the Figure 4. The filter used in the evaluation $z - z^{-1}$ is compared with the first-order filter $1 - \alpha z^{-1}$ for different values of α. Summarizing, the best performance is obtained by the second-order filter.

4 Conclusions

In this paper we have described three techniques for acoustic person identification in smart room environments. A baseline system based on a single microphone processing, SDM, has been described. Gaussian Mixture Model and a front-end based on Frequency Filtering has been used to perform the speaker recognition. To improve the mono-channel results, two multi-channel strategies are proposed. The first one based on a Delay and Sum algorithm to enhance the signal input and compensate the noise reverberations. The other one, based on a decision voting rule of three identical SDM systems.

The results show that the presented single distant microphone approach is well adapted to the conditions of the experiments. The use of D&S to enhance the signal has not show an improvement on the single channel results. The beamformed signal seems to lose some discriminative information that degrades the performance of the GMM classifier. However, the fusion of several single microphone decisions have really outperforms the SDM results in most of the train/test conditions.

Acknowledgements

This work has been partially supported by the EC-funded project CHIL (IST-2002 − 506909) and by the Spanish Government-funded project ACESCA (TIN2005 − 08852). Authors wish to thank Dusan Macho for the real time front-end implementation used in this work.

References

1. Casas, J., Stiefelhagen, R.: Multi-camera/multi-microphone system design for continuous room monitoring. In: CHIL Consortium Deliverable D4.1 (2005)
2. CLEAR-CONSORTIUM: Classification of Events, Activities and Relationships: Evaluation and Workshop (2007), http://www.clear-evaluation.org
3. Mostefa, D., et al.: CLEAR Evaluation Plan 07 v0.1 (2007), http://isl.ira.uka.de/clear07/?download=audio_id_2007_v0.1.pdf
4. Nadeu, C., Paches-Leal, P., Juang, B.H.: Filtering the time sequence of spectral parameters for speech recognition. In: Speech Communication, vol. 22, pp. 315–332 (1997)
5. Reynolds, D.A.: Robust text-independent speaker identification using Gaussian mixture speaker models. In: IEEE Transactions ASSP, vol. 3(1), pp. 72–83 (1995)
6. Flanagan, J., Johnson, J., Kahn, R., Elko, G.: Computer-steered microphone arrays for sound transduction in large rooms. In: ASAJ, vol. 78(5), pp. 1508–1518 (1985)
7. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: IEEE Transactions ASSP, vol. 28, pp. 357–366 (1980)
8. Nadeu, C., Macho, D., Hernando, J.: Time and Frequency Filtering of Filter-Bank Energies for Robust Speech Recognition. In: Speech Communication, vol. 34, pp. 93–114 (2001)
9. Furui, S.: Speaker independent isolated word recognition using dynamic features of speech spectrum. In: IEEE Transactions ASSP, vol. 34, pp. 52–59 (1986)
10. Nadeu, C., Hernando, J., Gorricho, M.: On the Decorrelation of filter-Bank Energies in Speech Recognition. In: EuroSpeech, vol. 20, p. 417 (1995)
11. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. In: IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 24(4), pp. 320–327 (1976)
12. Luque, J., Anguera, X., Temko, A., Hernando, J.: Speaker Diarization for Conference Room: The UPC RT 2007 Evaluation System. LNCS, vol. 4625, pp. 543–553. Springer, Heidelberg (2008)