

# 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, we describe the UPC person identification system submitted for the CLEAR'07 Evaluation. First we introduce the single microphone identification technique, as well as the comparison with the previous evaluation, and then we present the use of combined microphone inputs in two different approaches. The first one uses the input from several microphones to obtain a single enhanced signal through a delay and sum algorithm. The second one employs three identical algorithms which performs the identification from three different microphones, next a simple stage fusion, at the decision level, is in charge to obtain a unique score. The results presented shows that with a redundant identification and a simple decision fusion criterion we can improve the results of the single microphone approach. In the other side, the approach based on the beamformed signal is not well adapted to the task.

## 1 Introduction

During last years the interest for the identification of speakers in the surrounding of the smart environments has grown exponentially. The branch of the possibilities of the applications is huge, since the surveillance and access control of places up to the chance of offering a specific services depending the person identity. In a smart room, the typical situation is to have one or more cameras and several microphones which can gather relevant information to recognise, model and interpret human activity, behaviour and actions. The correct interpreting of all the information is the key which can lead us to a new concept of the sharing information in the interaction between the human and the computer, converting the two sides more independent from each one.

In this work we present the acoustic person identification technique and the obtained results in the CLEAR'07 Evaluation Campaign in the framework of the CHIL (Computers in the human interaction loop) project. This last evaluation follows the same criteria that the previous two evaluations, since it is a closed-set task, that means, all the possible speakers are known. The main goal of the CLEAR Identification Evaluation is the study of the degradation in function of the amount of speaker training data and the testing data, this last one used to perform the identification. In most of the real situations we do not have the

enough data to obtain an accurate estimation of the person model and as we have seen in the last evaluations, the performance degradation is a common feature of most of the systems presented when these are working in this situation. For instance, the systems show a big drop in the correct identification rates from the 5 second to the 1 seconds testing conditions.

Matched training and testing conditions and far-field data acquisition are assumed, as well as a limited amount of training data and no a priori knowledge about the room environment. In addition, in this CLEAR'07 Evaluation the multimicrophone recordings from the MarkIII array were provided to perform the acoustic identification contrasting with the last evaluation in which only one microphone was used in the testing stage. Therefore the using of multimicrophone data and their result comparison between the single microphone approach is other topic of studying in this evaluation.

This paper is organised as follows: In sections 2 an overview of the audio algorithms and techniques is given. The single distant microphone technique is described and two multimicrophone 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

In this section we describe the main features of the UPC acoustic speaker identification system. The three evaluated approaches shared the same characteristics about the parametrisation and statistical modelling, but differ in the use of the multichannel information. We first summarise the single distant microphone system (SDM) approach used in this work as a baseline system. Next we present a signal beamforming approach in which a set of channels is used to obtain an enhance signal through a delay and sum algorithm. Finally, we describe a fusion scheme which, through three identical SDM systems and a simple fusion of decisions, computes the person identity.

### 2.1 Single Distant Microphone System

The audio was analysed in frames of 30 milliseconds at intervals of 10 milliseconds and 16 kHz of sampling frequency. Each frame window is processed subtracting the mean amplitude from each sample, supposing the DC offset is constant throughout the waveform. A Hamming window was applied to each frame and a FFT computed. The FFT amplitudes were then averaged in 30 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. Their output represents the spectral magnitude in that filter-bank channel. Instead of the using Discrete Cosine Transform, such as in the Mel-Frequency Cepstral Coefficients (MFCC) procedure [1], the samples are parametrised using the Frequency Filtering (FF), see Eq. 1,

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

over the log of the filter-bank energies. Finally we obtain 30 FF coefficients. From each speech segment a vector of parameters is calculated. The choice of this kind of parameters is based on the fact that the using of the FF instead the classic MFCC has shown better results in both speech and speaker recognition [2]. This features have shown both computational efficiency and robustness against noise than the MFCC. We can find other interesting characteristics such as they are uncorrelated, have frequency meaning and are computationally simpler than MFCCs.

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  $\Delta$  and  $\Delta\text{-}\Delta$  coefficients are also used in this work. Note that the first coefficient of the FF output, which is proportional to the signal energy, is also employed to compute the model estimation as well as its velocity and acceleration parameters. Next, for each speaker that the system has to recognise, a model of the probability density function of the parameter vectors is estimated. These models are known as Gaussian Mixture Models (GMM) [3], which is a weighted sum of Gaussian distributions. A GMM of size 64 was used in this work. The parameters of the models were estimated from speech samples of the speakers using the well-known Baum-Welch algorithm. Given a collection of training vectors, maximum likelihood model parameters were estimated using the iterative Expectation-Maximisation (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 10 iterations were enough for parameter convergence in both training conditions and the same value was used for all the client models.

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

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

where  $s$  is the recognised speaker,  $L$  is the likelihood function from a linear combination of  $M$  unimodal Gaussian of dimension  $D$ , this expression can be found in the literature [4]. Therefore,  $L(\mathbf{O}|\lambda_j)$  is the likelihood that the vector  $\mathbf{O}$  has generated by the speaker of the model  $\lambda_j$ . The microphone number 4 was selected from the MarkIII array with the purpose of comparing with the last CLEAR Evaluation which the same one.

## 2.2 Delay-and-Sum Acoustic Beamforming

The delay and sum (D&S) beamforming technique [5] is a simple yet effective 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. The principle of operation of D&S can be seen in Figure 1.

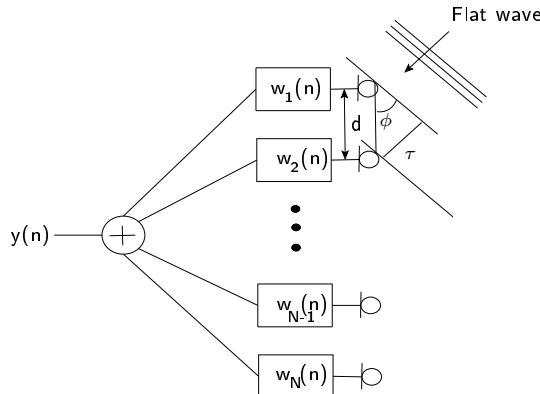


Fig. 1. Filter and sum algorithm block diagram

Given the signals captured by  $N$  microphones,  $x_i[n]$  with  $i = 0 \dots N - 1$  (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 using equation 3.

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

By adding together the aligned signals the usable speech adds together and the ambient noise (assuming it is random and has a similar probability function) will be reduced. Using D&S, according to [5], we can obtain up to a 3dB SNR improvement each time that we double the number of microphones. In order to estimate the TDOA between two segments from two microphones we used the generalised cross correlation with phase transform (GCC-PHAT) method ???. The TDOA for two microphones is estimated as:

$$\hat{d}_{PHAT}(f) = \arg \max_d \hat{R}_{PHAT}(d) \quad (4)$$

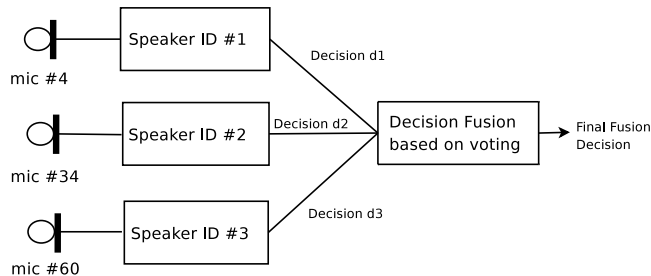
where  $\hat{R}_{PHAT}(d)$  is the inverse Fourier transform of  $G_{PHAT}(f)$ , the Fourier Transform of the estimated cross correlation phase. The maximum value of  $\hat{R}_{PHAT}(d)$  corresponds to the estimated TDOA.

In this work we have estimated the TDOA value using the whole sentence. In the testing stage the TDOA value is obtained as the maximum value of  $\hat{R}_{PHAT}(d)$ , obtaining this estimation from different window size depending of the duration of the testing sentence (1s/5s/10s/20s). In the training side, 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. The weighting factor  $W$  applied to each microphone is fixed to the inverse of the number of channels, since we suppose the microphones have the same response. A total of 20 microphones was used to perform the acoustic beamforming, selecting 1 of each 3 from the MarkIII 64 channels array. Once the TDOA values are estimated the beamformed signal is obtained, using Equation 3, in both testing and enrolment stages.

### 2.3 Multimicrophone Decision Fusion

In this approach we have implemented a redundant SDM system, such as described in Section 2.1 with three identical Speaker ID system working on three different microphone inputs. We have used the microphones number: 4, 34 and 60 which are applied independently to obtain a ID decision.



**Fig. 2.** In this Decision Fusion scheme we have used three microphones inputs. Next the individual decisions, a voting system is in charge to select the correct ID between the three identifications corresponding to each SDM system

Although the system applied to each microphone is the same, the three classifiers sometime are not agree about the identification of the segment data. It may be possible since the input signal differs between them due different incoming reverberation or other noises. In order to decide a sole ID in function of each classifiers output, a simple fusion of decisions is applied based on the following voting rule,

$$\begin{aligned}
 &\text{if } D_i \neq D_j \quad \forall i, j \neq i && \text{select the central microphone ID} && (5) \\
 &\text{if } D_i = D_j \quad \text{for some } i \neq j && \text{select } D_i
 \end{aligned}$$

where  $D_i$  is the decision of the classifier number  $i$ . Therefore in the case of three classifiers, we decide a ID if two of them are agree, and it decides the central microphone decision in the case all three classifier have not the same ID. The selection of the central microphone decision is motivated by its better single ID performance in our development experiments.

### 3 Experiments and Discussion

#### 3.1 Experimental set-up

A set of audiovisual 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. The recordings were done according to the "CHIL Room Setup" specification [6]. A complete description of the different recordings can be found in [7]. Data segments are far-field audio recordings taken from the above seminars. In order to evaluate how the duration of the training signals affects the performance of the system two training durations have been considered: 15 and 30 seconds. Test segments of different durations (1, 2, 5, 10 and 20 seconds) have been used during the algorithm development and testing phases. A total of 28 personal identities have been used in the recognition experiments.

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

**Table 1.** Number of segments for each test condition

For each seminar a 64 microphone channels, at 44.1 kHz and 16 bits/sample, from the MarkIII array were provided in contrast with the previous evaluation which was a monochannel evaluation. However, all processing were performed on 16 kHz. Each audio signal has been divided into segments which contain information of a sole speaker. These segments have been merged to form the final testing segments of 1, 5, 10 and 20 seconds (see Table 1) and training segments of 15 and 30 seconds. The metric used to benchmark the quality of the algorithms is the percentage of correctly recognised people from test segments. The silences longer than one second are removed from the data. That is the reason why we have not used a speech activity detection (SAD) in the front-end of our implementations.

#### 3.2 Results

In this section we summarise the results for the evaluation of the UPC acoustic modality in the CLEAR'07 Evaluation and examine the differences between the previous evaluation. Table 2 and 3 show the correct identification rate for both 2006 and 2007 CLEAR Speaker Identification Evaluation using the single microphone approach. In addition, we also can see the identification

rates obtained by the two multi-microphone implementations described on the previous section.

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, in general, are better as the segments length increases. The tables show that for the different test segment duration the recognition rate increases when more data is used to test the speaker models. Overall, using the 30 seconds training segments, an improvement of up to 6% in the recognition rate is obtained with respect to the case where 15 seconds segments are used.

Train A				
Duration	SDM'06	SDM'07	Fusion	D&S
1s	75.04 %	78.6 %	<b>79.6 %</b>	65.8 %
5s	89.29 %	<b>92.9 %</b>	92.2 %	85.7 %
10s	89.27 %	<b>96.0 %</b>	95.1 %	83.9 %
20s	88.20 %	<b>98.2 %</b>	97.3 %	91.1 %

**Table 2.** Percentage of correct identification for the UPC acoustic person identification system in the TRAIN A condition. The first column shows the duration of test segments in seconds. The table shows the rates obtained for the single microphone (SDM), Decision Fusion and Beamforming (D&S) systems. In addition, results from the single channel CLEAR'06 Evaluation are also provided.

Train B				
Duration	SDM'06	SDM'07	Fusion	D&S
1s	84.01 %	83.3 %	<b>85.6 %</b>	72.2 %
5s	97.08 %	95.3 %	<b>96.2 %</b>	89.5 %
10s	96.19 %	<b>98.7 %</b>	97.8 %	87.5 %
20s	97.19 %	<b>99.1 %</b>	<b>99.1 %</b>	92.9 %

**Table 3.** Percentage of correct identification for the UPC acoustic person identification system in the TRAIN B condition. The first column shows the duration of test segments in seconds. The table shows the rates obtained for the single microphone (SDM), Decision Fusion and Beamforming (D&S) systems. In addition, results from the single channel CLEAR'06 Evaluation are also provided.

On one hand, As we can see in the Tables 2 and 3 the Delay and Sum system is not well adapted to the task. The low performance of this implementation may be due to a not accurate estimation of the TDOA values. Other possibility could be the background noise and the reverberation effects of each partner room which could aid the GMM system to discriminate between recorded speakers

from different room setups. On the other hand, the decision fusion system seems, although the simple implementation, to exploit the redundant information from the multi-identification system. This technique achieves the better results in the 1s test conditions in both enrolment sets and, in general, in all the conditions of the train B set.

## 4 Conclusions

In this paper we have described three techniques for acoustic person identification in smart room environments. We have described a baseline system based on a single microphone processing. Gaussian Mixture Model of the distribution of the Frequency Filtering coefficients has been used to perform speaker recognition. To improve the obtained 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 an enhanced signal has not shown an improvement of the single channel results. The beamed signal seems to lose some necessary information that degrades the performance of the GMM classifier. However, we have improved the results of the SDM by the fusion of several single microphone decisions. The results show that this technique can provide an improvement of the recognition rate in some train/test conditions and further work will be done in this line.

## Acknowledgements

This work has been partially sponsored by the EC-funded project CHIL (IST-2002 – 506909) and by the Spanish Government-funded project ACESCA (TIN2005 – 08852).

## References

1. Davis S. B., M.P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: IEEE Transactions ASSP. Volume Vol. 28. (1980) pp. 357–366
2. Nadeu, C., Macho, D., Hernando, J.: Time and Frequency Filtering of Filter-Bank Energies for Robust Speech Recognition. In: Speech Communication. Volume 34. (2001) 93–114
3. Reynolds, D.A.: Robust text-independent speaker identification using Gaussian mixture speaker models. In: IEEE Transactions ASSP. Volume 3 NÅž 1. (1995) pp. 72–83
4. et al., F.B.: A Tutorial on Text-Independent Speaker Verification. In: EURASIP. (2004) vol.4, 430–451
5. J. Flanagan, J. Johnson, R.K., Elko, G.: Computer-steered microphone arrays for sound transduction in large rooms. In: Journal of the Acoustic Society of America. (1994)
6. J. Casas, R.S.: Multi-camera/multi-microphone system design for continuous room monitoring,. In: CHIL Consortium Deliverable D4.1. (2005)
7. et al., D.M.: CLEAR Evaluation Plan v1.1. In: <http://is1.ira.uka.de/~nickel/clear/downloads/chil-clear-v1.1-2006-02-21.pdf>. (2006)