Speaker Diarization for Conference Room: The UPC RT07s Evaluation System

Jordi Luque[1], Xavier Anguera[2], Andrey Temko[1], and Javier Hernando[1]

[1] Technical University of Catalonia (UPC),
Jordi Girona, 1-3 D5, 08034 Barcelona, Spain,
luque@tsc.upc.edu
[2] Multilinguism group, Telefónica I+D,
08021 Barcelona, Spain

Abstract. In this paper the authors present the UPC speaker diarization system for the NIST Rich Transcription Evaluation (RT07s) [1] conducted on the conference environment. The presented system is based on the ICSI RT06s system, which employs agglomerative clustering with a modified Bayesian Criterion (BIC) measure to decide which pairs of clusters to merge and to determine when to stop merging clusters [2]. This is the first participation of the UPC in the RT Speaker Diarization Evaluation and the purpose of this work has been the consolidation of a baseline system which can be used in the future for further research in the field of diarization. We have introduced, as prior modules before the diarization system, an Speech/Non-Speech detection module based on a Support Vector Machine from UPC and a Wiener Filtering from an implementation of the QIO front-end. In the speech parameterization a Frequency Filtering (FF) of the filter-bank energies is applied instead the classical Discrete Cosine Transform in the Mel-Cepstrum analysis. In addition, it is introduced a small changes in the complexity selection algorithm and a new post-processing technique which process the shortest clusters at the end of each Viterbi segmentation.

1 Introduction

Audio segmentation, sometimes referred to as acoustic change detection, consists of exploring an audio file to find acoustically homogeneous segments, detecting any change of speaker, background or channel conditions. It is a pattern recognition problem, since it strives to find the most likely categorization of a sequence of acoustic observations. Audio segmentation becomes useful as a preprocessing step in order to transcribe the speech content in broadcast news and meetings, because regions of different nature can be handled in a different way.

There are two basic approaches to this problem: (1) *model-based* segmentation [3], which estimates different acoustic models for a closed set of acoustic classes (e.g. noise, music, speech, etc.) and classifies the audio stream by

finding the most likely sequence of models; and (2) *metric-based* segmentation [4], which defines some metric to compare the spectral statistics at both sides of successive points of the audio signal, and hypothesizes those boundaries whose metric values exceed a given threshold. The first approach requires the availability of enough training data to estimate the models of acoustic classes and does not generalize to unseen conditions. The second approach, sometimes referred as *blind* (unsupervised) segmentation, does not suffer from these limitations, but its performance depends highly on the metric and the threshold. Various metrics have been proposed in the literature. The most cited are the *Generalized Likelihood Ratio* (GLR) [5] and the *Bayesian Information Criterion* (BIC) [4].

The Diarization task assume no prior knowledge about the speakers or how many people participate in the meeting. In order to get acquainted with the problem, the data and the evaluation methodology, we have taken as a baseline a simplified version of the International Computer Science Institute (ICSI) RT06s system as presented in [2]. Our submission still uses the multi-channel and agglomerative clustering capabilities from ICSI's software while using our own Speech Activity Detection (SAD) algorithm, parameterization and avoiding the use of several algorithms in order to make the system more lightweight. Hence we have used an approach which performs the clustering through a modified BIC measure to decide which pairs of clusters to merge and to determine when to stop merging clusters, as in [2].

In addition, some novelties to the diarization system are studied. The use of the Frequency Filtering (FF) parameters instead the classical Mel Frequency Cepstral Coefficients (MFCCs) has been introduced in the speech parameterization. Other of them, a post-processing module is applied after each Viterbi decoding. It looks for orphan speaker segments with small duration and splits them between the adjacent segments. Other new feature is a small modification to the cluster complexity algorithm. It avoids the creation of very small clusters, which do not alter the real system outcome greatly but do pose a burden on execution time.

The following sections give a brief overview of the diarization system focusing in the novelties introduced. Finally, the results section provides the Diarization Error (DER) obtained by the system in the NIST RT07S Evaluation and some comments.

2 System description

The input signal from each one of the multiple distant microphones (mdm) channels, if they are available, is first **Wiener filtered** using the implementation from the QIO front-end [6]. These channels are then fed into the **Beamforming** code implemented by ICSI [7] in order to obtain a single enhanced channel to be further processed. Such output channel is analyzed by the Speech Activity Detector (SAD) from UPC [8] in order to obtain the Speech segments to be fed into the clustering algorithm. The Non-Speech segments are ignored from further

processing. The enhanced speech data is parameterized using 30 Frequency Filtering (FF) features as described in [9] and fed into an evolution of ICSI's speaker agglomerative clustering system [2].

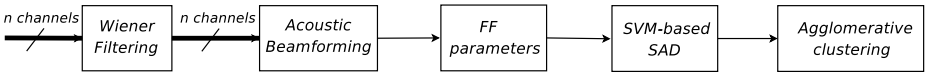

Fig. 1. *Brief scheme of the UPC implementation of the RT'07 diarization system*

2.1 Wiener Filtering

We used the noise reduction implementation from the QIO front-end [6]. The Wiener filter noise reduction technique was applied on each channel input waveform. That depends on a noise estimate made over frames judged to be Non-Speech. Despite the SAD used in the next stages of the diarization system, in this phase we used the procedure from the QIO front-end: The noise estimate is initialized from the beginning of each utterance, assuming each sentence starts with a period of Non-Speech, and updated using later frames of the utterance decided to be Non-Speech based on an energy threshold.

2.2 Acoustic Beamforming

The Delay-and-Sum (D&S) technique [10] is one of the simplest beamforming techniques but still gives a very good performance. It is based on the fact that applying different phase weights to the input channels the main lobe of the directivity pattern can be steered to a desired location, where the acoustic input comes from. It differs from the simpler D&S beamformer in that an independent weight is applied to each of the channels before summing them. The principle of operation of D&S can be seen in Figure 2.

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 time delay of arrival due the different positions of the microphones with regard to the source. So if we estimate the time τ, see Figure 2, 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 using Equation (1).

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

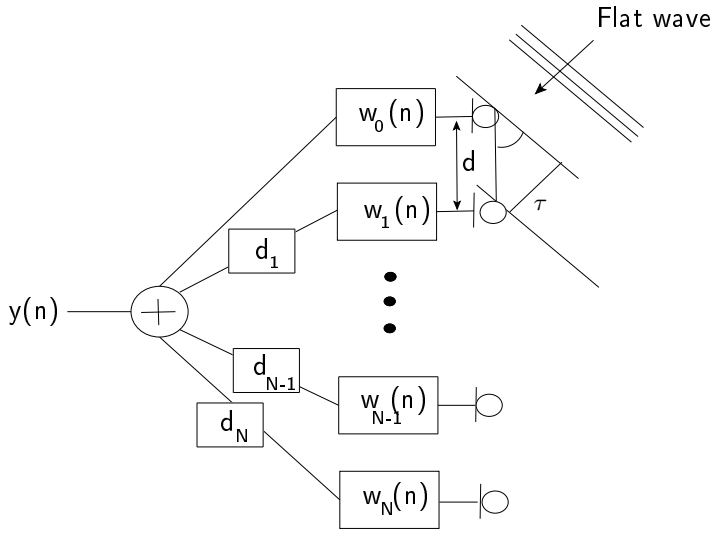

Fig. 2. *Filter and Sum algorithm block diagram*

The same technique is applied in both training and testing speech leading to matched conditions in the identification. 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 [10], 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 [11]. Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is defined as:

$$\hat{G}_{PHAT_{ij}}(f) = \frac{X_i(f)\big[X_j(f)\big]^*}{\big|X_i(f)\big[X_j(f)\big]^*\big|} \tag{2}$$

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:

$$\hat{d}_{PHAT_{ij}} = \arg\max_d \hat{R}_{PHAT}(d_{ij}) \tag{3}$$

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.

In this work we have estimated the TDOA value using a window of 500 ms. at rate of 250 ms. applied on the wiener filtered channels. During the development some experiments were performed with different sizes and shifts of window, but

we did not find any improvement in the overall DER error. The weighting factor W applied to each microphone is computed depending the cross correlation between each channel and the reference channel.

2.3 SVM-based Speech Activity Detection

The SAD module used in this work is based on SVM classifier [12]. The developed system showed a good performance in the last RT SAD Evaluations [8], hence we have chosen this SAD implementation due to the fact it is adapted to NIST metric about speech activity detection since it penalizes more the Speech class than the Non-Speech class.

The usual training algorithm of the SVM classifier was enhanced in order to cope with that problem of dataset reduction, proposing a fast algorithm based on Proximal SVM (PSVM). Besides that, the SVM learning process was adjusted in order to take into account the specific characteristics of the metric used in the NIST Rich Transcription (RT) evaluations. The resulting SVM SAD system was tested with the RT06 data and it showed better scores than the GMM-based system which ranked among the best systems in the RT06 evaluation [8].

A set of several hundred of thousand of examples is a usual amount of data for classical audio and speech processing techniques that involve GMM. However, it is an enormous number of feature vectors to be used for a usual SVM training process and hardly makes such training feasible in practice. Alternative methods should be effectively applied to reduce the amount of data.

Proximal Support Vector Machine (PSVM) has been recently introduced in [13] as a result of the substitution of the inequality constraint of a classical SVM $y_i(wx_i + b) \geq 1$ by the equality constraint $y_i(wx_i + b) = 1$, where y_i stands for a label of a vector x_i, w is the norm of the separating hyperplane H_0, and b is the scalar bias of the hyperplane H_0. This simple modification significantly changes the nature of the optimization problem. Unlike conventional SVM, PSVM solves a single square system of linear equations and thus it is very fast to train. As a consequence, it turns out that it is possible to obtain an explicit exact solution to the optimization problem [13].

Figure 3 shows a geometrical interpretation of the change. H_{-1} and H_1 planes do not bound the negatively- and the positively-labeled data anymore, but can be viewed as *proximal* planes around which the points of each class are clustered and between which the separating hyperplane H_0 lies. In the nonlinear case of PSVM (we use a Gaussian kernel) the concept of Support Vectors (SVs) (Figure 3, in gray) disappears as the separating hyperplane depends on all data. In that way, all training data must be preserved for the testing stage.

The proposed algorithm of dataset reduction consists of the following steps:

- Step 1. Divide all the data into chunks of 1000 samples per chunk
- Step 2. Train a PSVM on each chunk performing 5-fold cross-validation (CV) to obtain the optimal kernel parameter and the C parameter that controls the training error

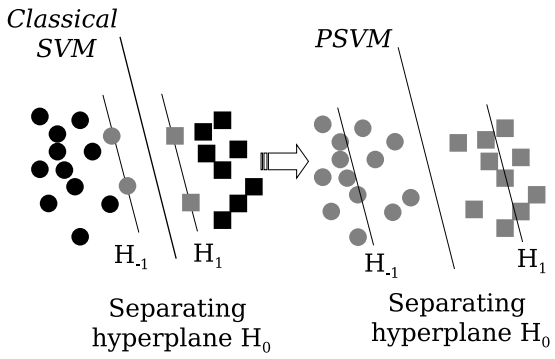

Fig. 3. *Proximal Support Vector Machine based SVM.*

- Step 3. Apply an appropriate threshold to select a pre-defined number of chunks with the highest CV accuracy
- Step4. Train a classical SVM on the amount of data selected in Step 3

The proposed approach is in fact similar to Vector Quantization (VQ) used for dataset reduction for SVM in [14]. With Step 2 some kind of clustering is performed, and Step 3 chooses the data that corresponds to the most separable clusters. However, unlike VQ, SVs, which are obtained with the proposed algorithm in Step 4, are taken from the initial data. Besides, additional homogeneity is achieved because the PSVM data clustering is performed in the transformed feature spaces with the transformation functions that correspond to the Gaussian kernel and the same kernel type is applied to the chosen data in Step 4.

The second modification makes use of the knowledge of the specific NIST metric during the training phase. The NIST metric depends strongly on the prior distribution of Speech and Non-Speech in the test database. For example, a system that achieves a 5% error rate at Speech portions and a 5% error rate at Non-Speech portions, would result in very different NIST error rates for test databases with different proportion of Speech and Non-Speech segments; in the case of 90-to-10% ratio of Speech-to-Non-Speech the NIST error rate is 5.6%, while in the case of 50-to-50% ratio it is 10%. For this reason, if we want to improve the NIST scores we should penalize the errors from the Speech class more than those from the Non-Speech class. That is possible for a discriminative classifier as SVM in the training stage by introducing different costs for the two classes through the different generalization parameters C_- and C_+. In that way, the separating hyperplane H_0 will no longer lie exactly in the middle of the H_{-1} and H_1 hyperplane (Figure 3). It is worth to mention that favouring a class in the testing stage (after the classifier is trained) could still be done for SVM through the bias b of the separating hyperplane.

2.4 Speech Parameterization

The speech parameterization is based on a short-term estimation of the spectrum energy in several sub-bands. The beamformed channel was analyzed 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. 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.

The squeme we present follow the classical procedure used to obtain the Mel-Frequency Cepstral Coefficients (MFCC), however in this approach, instead of the using Discrete Cosine Transform, such as in the MFCC procedure [15] log filter.bank energies are filtered by a linear and second order filter. This technique was called Frequency Filtering (FF) [9]. The filter $H(z) = z - z^{-1}$ have been used in this work and it's applied over the log of the filter-bank energies. The shape of this filter allow a best classification due it emphasizes regions of the spectrum with high speaker information yielding more discriminative information. This parameters have show a good results in the last CLEAR Evaluation Campaign in the acoustic person identification task [16].

A total of 30 FF coefficients had used in this work and no Δ or $\Delta - \Delta$ parameters. The choice of this kind of parameters is based on the fact that the using of the FF instead the classic MFCC has shown the best results in both speech and speaker recognition [17]. This features have shown both computational efficiency and robustness against noise than the MFCC. In addition, regarding in the frequency domain imply they have frequency meaning which permits the use of frequency techniques as masking, noise subtraction, etc. We can find other interesting characteristics such as they are uncorrelated, computationally simpler than MFCCs and it does not decrease clean speech recognition results [18]. Summarizing, the FF filter can be seen as a liftering operation performed in the spectral domain equalizing the variance of cepstral coefficients.

2.5 Improvements in agglomerative clustering

Thequeme we present follow the classical procedure used to obtain the Mel-Frequency Cepstral Coefficients (MFCC), however in this approach, instead approach is based on a iterative segmentation by an ergodic Hidden Markov Model (HMM), which models the acoustic data and their temporal evolution. The system starts with a homogeneous splitting of the data among an initial number of cluster equal to the initial number of states. Next the Viterbi decoding, it merges the pair of cluster more acoustically similar by a modified version of BIC [19]. The BIC measure also handles the stop criterion which occurs if the remaining clusters are below a threshold in the likelihood function. In the end-iteration each remaining state is taken to represent a different speaker.

Changes to the diarization system from ICSI are oriented towards decreasing the runtime of the system while maintaining as much as possible the performance

from the original. For instance, in this version it does not use delays as features, does not perform any kind of purification to the clusters and uses linear initialization by splitting evenly all data among the number of determined initial clusters.

In addition, novelties to the diarization system are a post-processing module that looks for orphan speaker segments with small duration and splits them between both adjacent segments and a small modification to the cluster complexity algorithm to avoid the creation of very small clusters, which do not alter the real system outcome greatly but do pose a burden on execution time.

The cluster complexity modification allow drop off small clusters which are modelled by a few Gaussians. The class pruned does not take part in the following segmentations and after the next segmentation step its data is splitted among the remaining classes.

At the end of each segmentation, the final post-processing of the boundaries analyzes whose shortest segments normally associated to false alarms in this kind of tracking implementation. All those segments with duration small than 1.1*MD (Minimum Duration) are processed through a sliding window. From the pre-boundary up to the post-boundary all data inside the window are evaluated using the model of the previous, current and posterior cluster and the new boundary is chosen depending the maximum computed likelihood. Once the last iteration is completed, the system reach the stop criteria and next the last post-processing of the boundaries, the final hypothesis is obtained.

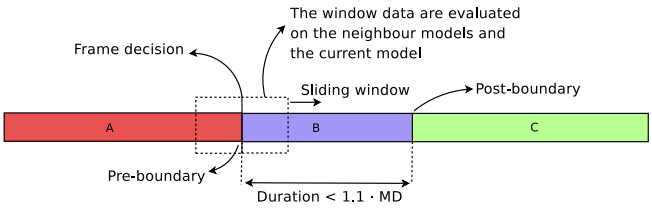

Fig. 4. *The final stage of the algorithm consists in a post-processing of the short segment boundaries at the final of each segmentation. A sliding windows are applied on the shortest clusters in order to decide with more accuracy the real boundaries*

3 Experiments and results

This section summarizes the results for the evaluation of the UPC diarization system. It examines the differences between the two SAD depending systems as well as the improvement achieved by the mdm systems compared to the sdm approach.

The Table 1 shows the performance of the SAD module in the different Rich Transcription Evaluations conducted in the surrounding of the conference room environment.

SAD SVM-based				
RT'05 sdm	RT'06 sdm	RT'07 sdm	RT'07 mdm-softsad	RT'07 mdm-hardsad
8.03 %	4.88 %	7.03 %	5.39 %	4.72 %

Table 1. SAD error results in the previous RT Evaluation Conference data condition.

The difference between the two mdm systems submitted is the SAD behavior. The bias b of the separating hyperplane, see section 2.3, is chosen according to the speaker time error and false alarms produced in the Non-Speech segments of development data. The weighting which controls the decision boundary between the Speech/Son-Speech classes is chosen according speaker time error and false alarms of Non-Speech using the NIST RT06s Evaluation data. That is the same fashion than the development of the overall diarization system. The conference NIST RT06s evaluation was used to perform all the development experiments for the RT07s.

The Tables 2 and 3 shows the results obtained by the UPC implementation in the RT07s. As we can see, the result from the single distant condition is improved in the multichannel approach. Other interesting feature is the behavior of the diarization system in function of the SAD performance. The system seems to behave in a similar fashion in spite of the differences of the SAD applied. However, more and accurate experiments must be done in this line trying to find the tradeoff between the speech false alarms and the diarization performance.

Overlap SPKR Error, **Primary Metric**		
sdm	mdm-softsad	mdm-hardsad
27.72 %	22.70 %	22.59 %

Table 2. RT07s Diarization error results of the UPC implementation using the Primary Metric of NIST which considers overlapping of speaker segments

Non-Overlap SPKR Error		
sdm	mdm-softsad	mdm-hardsad
25.06 %	19.65 %	19.75 %

Table 3. RT07s Diarization error results of the UPC implementation without considering overlapping of speaker segments

Finally, in the Table 4 we can see the RT07s DER per show of the *mdm-softsad* system as well as some experiments performed after the Evaluation. We can observe a high variance between the DER errors from different shows, motivated by the difficulty to tune all the parameters using the RT06s data, around 4 hours of speech. The *mdm-noE* system differs from the *mdm-softsad* only in the number of FF parameters, it uses a vector size of 28. This system

does not use the first and last coefficients of the FF. Note that the first and the last coefficients of the FF output of each frame contain absolute energy [20], so they can carry much noise. The last system in the Table, the *mdm-nocomplex* does not implements the modification of the complexity algorithm, it means, no pruning of the small clusters are done.

On the one hand, as we can note in Table 4, the use of the lateral-band coefficients performs badly in the diarization system and, in overall, it is better do not include this features in the speaker modelling. On the other hand, the pruning of small clusters on the complexity algorithm significantly affects the diarization error, over a 5 % of DER fall down by using the complexity algorithm modification, see Table 4 instead the original one from ICSI. Some experiments during the development showed a best behavior of this technique and it could be interesting to find the minimum cluster complexity out to decide the pruning as a tradeoff between the DER degradation and the runtime of the system.

	Overlap SPKR Error		
show	mdm-softsad	mdm-noE	mdm-nocomplex
CMU_20061115-1030	57.58 %	39.68 %	23.51 %
CMU_20061115-1530	11.46 %	12.64 %	15.12 %
EDI_20061113-1500	24.44 %	24.53 %	31 %
EDI_20061114-1500	17.97 %	15.39 %	17.16 %
NIST_20051104-1515	11.16 %	11.39 %	11.23 %
NIST_20060216-1347	5.62 %	11.4 %	10.77 %
VT_20050408-1500	7.13 %	6.9 %	7.44 %
VT_20050425-1000	49.02 %	34.3 %	28.66 %
DER global	**22.70** %	**19.36** %	**17.83** %

Table 4. RT07s Diarization error per show of the (*mdm-softsad) system. In addition some experiments posterior to the Evaluation are showed, one of them without using the first and last coefficients of the FF and the other one, without the modification of the complexity algorithm.*

4 Conclusions

In this work the authors have presented the UPC Diarization system and the results obtained in the NIST RT07s Diarization Evaluation on Conference room data. We have described and implementation of an agglomerative clustering approach based on a software from the ICSI. In addition some novelties are introduced in the diarization system. A Speech/Non-Speech detection module based on a Support Vector Machine is studied. In the speech parameterization the using of Frequency Filtering coefficients is introduced and minor modifications to the complexity selection algorithm and a new post-processing technique are tested looking for a runtime reduction while maintiaining as much as possible the performance of the system. The results

obtained in the RT07s show that the fine tuning of the SAD seems not affect significativily the DER of the global system. In addition, in the mdm approach, the DER achieved outperforms the results from the sdm algorithm in all show conditions. Therefore, the using of a simple delay-and-sum algorithm to enhace the signal aids the system to obtain a better clustering. Finally, the main goal of the UPC evaluation is achieved and a diarization system as baseline system for further development and research have been implemented with promising results.

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. NIST: Rich transcription meeting recognition evaluation plan. In: RT-07s. (2007)
2. Anguera, X., Wooters, C., Hernando, J.: Robust speaker diarization for meetings: Icsi rt06s evaluation system. In: ICSLP. (2006)
3. Gauvain, J., Lamel, L., Adda, G.: Partitioning and transcription of broadcast news data. In: ICSLP. (1998) 1335–1338
4. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: DARPA BNTU Workshop. (1998)
5. Gish, H., Siu, M., Rohlicek, R.: Segregation of speakers for speech recognition and speaker identification. In: ICASSP. (1991)
6. A. Adami, e.a.: Qualcomm-icsi-cgi features for asr. In: ICSLP. (2002) 21–24
7. Anguera, X.: The acoustic robust beamforming toolkit. (2005)
8. Temko, A., Macho, D., Nadeu, C.: Enhanced SVM Training for Robust Speech Activity Detection. In: Proc. ICCASP. (2007)
9. Nadeu, C., Paches-Leal, P., Juang, B.H.: Filtering the time sequence of spectral parameters for speech recognition. In: Speech Communication. Volume 22. (1997) 315–332
10. Flanagan, J., Johnson, J., Kahn, R., Elko, G.: Computer-steered microphone arrays for sound transduction in large rooms. In: ASAJ. Volume 78, No. 5. (1985) 1508–1518
11. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. In: IEEE Transactions on Acoustic, Speech and Signal Processing. Volume 24, No. 4. (1976) 320–327
12. Schölkopf, B., Smola, A.: Learning with Kernels. In: MIT Press, Cambridge, MA. (2002)
13. Fung, G., Mangasarian, O.: Proximal Support Vector Machine Classifiers. In: Proc. KDDM. (2001) 77–86
14. Lebrun, G., Charrier, C., Cardot, H.: SVM Training Time Reduction using Vector Quantization. In: Proc. ICPR. (2004) 160–163
15. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In: IEEE Transactions ASSP. Volume No. 28. (1980) 357–366

16. Luque, J., Hernando, J.: Robust Speaker Identification for Meetings: UPC CLEAR-07 Meeting Room Evaluation System. In: The same book. (2007)
17. 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
18. Macho, D., Nadeu, C.: On the interaction between time and frequency filterinf of speech parameters for robust speech recognition. In: ICSLP. (1999) paper 1137
19. Anguera, X., Hernando, J., Anguita, J.: Xbic: nueva medida para segmentación de locutor hacia el indexado automático de la señal de voz. In: JTH. (2004) 237–242
20. Nadeu, C., Hernando, J., Gorricho, M.: On the Decorrelation of filter-Bank Energies in Speech Recognition. In: EuroSpeech. Volume 20. (1995) 417