

SPEAKER RECOGNITION USING FREQUENCY FILTERED SPECTRAL ENERGIES

Javier Hernando

TALP Research Center,
TSC Dept., UPC, Barcelona, Spain
javier@gps.tsc.upc.es

ABSTRACT

The spectral parameters that result from filtering the frequency sequence of log mel-scaled filter-bank energies with a simple first or second order FIR filter have proved to be an efficient speech representation in terms of both speech recognition rate and computational load. Recently, the authors have shown that this frequency filtering can approximately equalize the cepstrum variance enhancing the oscillations of the spectral envelope curve that are most effective for discrimination between speakers. Even better speaker identification results than using mel-cepstrum have been obtained on the TIMIT database, especially when white noise was added. On the other hand, the hybridization of both linear prediction and filter-bank spectral analysis using either cepstral transformation or the alternative frequency filtering has been explored for speaker verification. The combination of hybrid spectral analysis and frequency filtering, that had shown to be able to outperform the conventional techniques in clean and noisy word recognition, has yield good text-dependent speaker verification results on the new speaker-oriented telephone-line POLYCOST database.

1. INTRODUCTION

In current speaker recognition systems, the short-time spectral envelope of every speech frame is usually represented by a set of the Fourier series coefficients of its logarithm, i.e. the cepstral coefficients $C(m)$, $1 \leq m \leq M$. These parameters are by far the most prevalent representations of speech signal and contain a high degree of speaker specificity [1]. They usually come either from a linear prediction (LP) analysis -LP-cepstrum-, or from a set of mel-scaled log filter-bank (FB) energies -mel-cepstrum. Unfortunately, there are few comparative studies about the relative robustness to

noise and distortions of mel-cepstrum with respect to LP-cepstrum.

Recently, the authors have considered a unified parameterization scheme for speech recognition that combines both LP and FB analysis [2]. It has been shown that an appropriate hybridization of both LP and FB approaches is capable of improving recognition results for both noisy and clean digit recognition.

On the other hand, we may wonder if the cepstral coefficients are the best way of representing the speech spectral envelope. The sequence of cepstral coefficients $C(m)$ is a quasi-uncorrelated and compact representation of speech spectra. The quefrequency sequence $C(m)$ is always windowed before entering a distance or probability computation in the pattern matching stage of the recognition process. That window eliminates the cepstral coefficients beyond a quefrequency M . And, for some type of recognition systems, it also appropriately weights the remaining coefficients [1] [3] [4] [5] in order to increase the discrimination capability of the system.

However, cepstral coefficients have at least three disadvantages: 1) they do not possess a clear and useful physical meaning as log FB energies have; 2) they require a linear transformation from either log FB energies or the LPC coefficients; and 3) in continuous observation Gaussian density HMM with diagonal covariance matrices, the shape of the cepstral window has no effect so that only its length, i.e. the number of parameters M , is a control variable.

In order to try to overcome those disadvantages, the authors have recently proposed an alternative spectral representation that result from filtering the frequency sequence of log energies with a simple FIR filter of order 1 or 2 for both speech [6] and speaker [7] [8] recognition. In the last works, it is shown that this frequency filtering can approximately equalize the cepstrum variance enhancing the oscillations of the spectral envelope curve that are most effective for discrimination between speakers.

The aim of this paper is to gain some perspective of the merit of the hybridization of both LP and FB spectral analysis (section 2) and the novel frequency filtering technique (section 4) in speaker recognition. It is shown

that frequency filtering produces both desired effects, decorrelation and discrimination (section 3), in one step. Using frequency filtering of log FB energies, even better results than using mel-cepstrum have been observed in text-independent speaker identification experiments on the TIMIT database, especially when white noise is added (section 5). Furthermore, the combination of both frequency filtering and hybrid spectral analysis techniques had shown to be able to outperform the conventional LP and mel-cepstrum in clean and noisy word recognition [2]. In this way, text-dependent speaker verification experiments on the new speaker-oriented telephone-line POLYCOST database have been performed and better results than using conventional LP and mel-cepstrum have been observed (section 6).

2. HYBRID SPECTRAL ANALYSIS

The strength of LP method arises from its close relationship to the digital model of speech production. So an appropriate deconvolution between vocal tract response and glottal excitation can be expected from it.

LP is a full-band approach to spectrum modeling. Conversely, the filter-bank (FB) approach removes pitch information and reduces estimation variance (error) by integrating the periodogram (the squared value of the DFT samples) in frequency bands. The FB approach separately models the spectral power for each band, and it offers the possibility of easily distributing the position of the bands in the frequency axis (a mel scale is employed in the so-called mel-cepstrum) and defining their width and shape in any desired way, to take advantage of the perception properties of the human auditory system. This sub-band working mode also has several advantages derived from the frequency localization of the parameters. For example, if the SNR of each band is known, it can be used in straightforward ways: noise subtraction, noise masking,...

The combination of LP and FB analysis may yield improved spectral parameters. One possible approach is to apply FB analysis on the signal prior to LP analysis [9] [10]. It will be referred to as FB-LP and it is computed similarly to the PLP coefficients [9], but using a higher order LP analysis without perceptual weighting and amplitude compression. An alternative approach is to use LP analysis followed by FB analysis (it will be referred to as LP-FB).

Both conventional LP and mel-cepstrum parameterizations and the cepstrum representations corresponding to the two new hybrid FB-LP and LP-FB can be encompassed in a unified parameterization scheme [2], that can lead to other novel speech parameterization techniques.

3. DECORRELATION AND DISCRIMINATION

HMM are mostly employed with diagonal covariance matrices. In that case, they implicitly assume uncorrelated spectral parameters. That is true for the Gaussian pdf of continuous density HMM (CDHMM) and semicontinuous density HMM (SCHMM), and also for the Mahalanobis distance of discrete HMM. Conversely, the frequency sequence of log FB energies $S(k)$ is strongly correlated. The usual mel-cepstrum is a way of obtaining from $S(k)$ an almost uncorrelated set of parameters. Actually, by approximating the random process $S(k)$ with a first-order Markov model, it follows that the discrete cosine transform is almost equivalent to the Karhunen-Loève transform.

Decorrelation is thus a desired property for the sets of spectral parameters due to the particular way they are used in our current recognition systems. And also because decorrelation may provide a less redundant representation. Nevertheless, what is really relevant to the own classification process is the discrimination capacity of those parameters.

It is a known fact that the variance of $C(m)$ decreases along the axis m [4]. Figure 1 shows an estimation of this variance for the TIMIT database using $Q=20$ mel-scaled frequency bands. Note the zero value corresponding to zero quefrency, which is caused by the subtraction of the average $S(k)$ value [11].

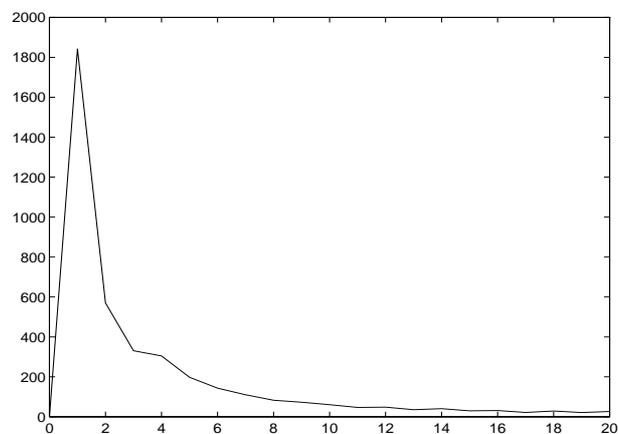


Figure 1. Variance of the cepstral coefficients for the TIMIT database.

Thus, the low quefrecies m will generally dominate the probability or distance computations in the classifier. We may ask whether this is the best we can do or a proper global variance equalization of $C(m)$ could help to increase recognition performance, much like it occurs in speech recognition [6]. Let us note that there exists a close relationship between equalization of the variance of $C(m)$ at low quefrecies and decorrelation of $S(k)$.

However, a flat variance may not be the most adequate goal for recognition purposes. For example, when the frequency interval between bands is not large enough, that equalization gives too much weight to the estimation noise carried out by $C(m)$ for high frequencies. Another reason for not flattening it completely can be the presence of the acoustic channel characteristics or broad-band additive noise, which may require a stronger attenuation of the lowest frequencies.

A possible measure of the discrimination capacity of each cepstral coefficient can be the ratio between its inter-speaker and global variances. Figure 2 shows an estimation of this ratio for the TIMIT database using $Q=20$ frequency bands.

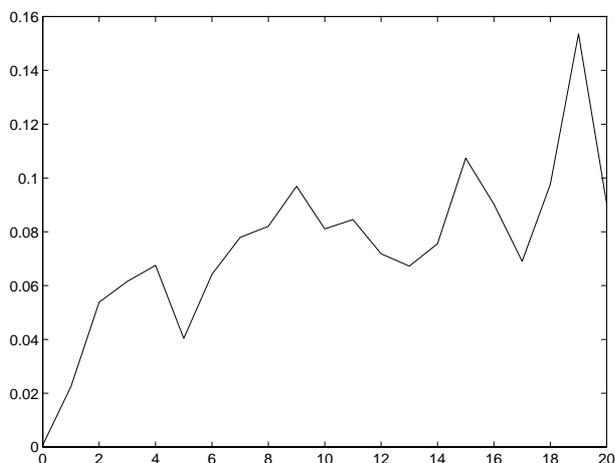


Figure 2. Estimation of the ratio between inter-speaker and global cepstrum variances for the TIMIT database

As it can be seen in Figure 2, the dynamic range of this ratio sequence is smaller than that of the global variance shown in Figure 1. This fact suggests that an approximate equalization of the variance can help to increase the discrimination capability of the cepstral sequence, at least for clean speech.

On the other hand, Figure 2 shows a slight increasing tilt along the frequency index m . This fact leads to think that the most discriminative information is located in the higher frequencies, i.e. in the fast alternation of peaks and valleys of the spectral curve; and it is not in the lowest frequencies, i.e. in the spectral tilt. Actually, most speaker recognition systems do use a higher number of cepstral parameters than speech recognizers. Even it could be convenient to slightly over-emphasize the higher frequencies.

Cepstral liftering (weighting on m) has been the usual way to compensate for the excessive weight of the lowest m terms in both speech and speaker recognition systems. In this case, two steps are needed for obtaining the final parameters from the log FB energies: 1) a linear transformation (discrete cosine transform), that

significantly decorrelates the sequence of parameters, and 2) a weighting (liftering) of the cepstral coefficients. Furthermore, in continuous observation Gaussian density HMM with diagonal covariance matrices, the shape of the cepstral window has no effect due to the intrinsic variance normalization of the Gaussian pdf.

In recent papers, in order to try to overcome those disadvantages and to have parameters that possess frequency meaning, an alternative to the use of cepstral parameters was introduced for speech [6] and speaker [7] [8] recognition. It consists in a simple linear processing in the log energy domain. The transformation of the sequence of log FB energies to cepstral coefficients is avoided by performing a filtering of that sequence, which we hereafter will call frequency filtering to denote that the convolution is performed on the frequency domain.

This frequency filtering produces both effects, decorrelation and discrimination, in only one step using a simple first or second order FIR filter. Moreover, frequency filtering is able to produce a cepstral weighting in an implicit way in continuous observation Gaussian density HMM with diagonal covariance matrices.

4. FREQUENCY FILTERING

We aim to perform an approximate equalization of the variance of the cepstral coefficients by filtering the frequency sequence of log energies. Since this filtering is implemented as a circular convolution with the sequence $h(k)$, the cepstral coefficients are multiplied (weighted) by the DFT of $h(k)$, here denoted by $H(m)$.

First of all, since in the usual mel-scaled FB there are not any filters centered at frequencies $\omega=0$ and $\omega=\pi$, a zero is appended at both ends of the sequence, i.e. $S(0)=S(Q+1)=0$, to represent the low energy contained at those extreme bands. Then, according to the usual practice [11], in every frame, the average value of the even sequence $S(k)$ over index k is subtracted.

After that, $S(k)$ is circularly convoluted with $h(k)$ to obtain a filtered sequence. Since only the values of the filtered sequence between $k=1$ and $k=Q$ are used as observations in the recognition system, we can employ the shortest $h(k)$, i.e. a length 2, with no interference of the symmetric $S(k)$, $k=-1, \dots, -Q$, samples in the computation of the used segment of the filtered sequence. In this way, we can refer to the process as an actual linear filtering, with $h(k)$ being the impulse response.

A first-order FIR filter that maximally equalizes the variance of the cepstral coefficients can be easily obtained by a least-squares modeling in the following way. Firstly, the variance is estimated by averaging over all the frames of a given database. Then, after performing an inverse DFT, the quotient r between the values of the resulting sequence –the variance of $S(k)$ – at index 1 and

index 0 is computed. Thus, the first-order FIR filter that maximally flattens the variance will be $H(z)=1-rz^{-1}$. Figure 3 shows the product of the cepstrum global variance corresponding to the TIMIT database using $Q=20$ mel-scaled frequency bands -shown in Figure 1- by the magnitude of the sampled filter response $H(m)$, that was computed following the above procedure. The resulting value of r is 0.75. As it can be seen, the cepstrum variance tilt has been approximately equalized.

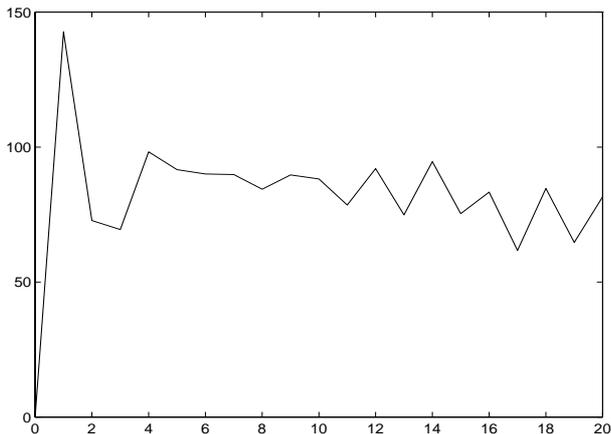


Figure 3. Cepstrum variance equalized by the sampled filter response $H(m)$.

However, the frequency filtering representation may improve its performance if this frequency filter is empirically optimized, perhaps taking into account the slight increasing tilt along the axis m of the ratio between inter-speaker and global cepstrum variances shown in Figure 2.

The spectral parameters that result from filtering the frequency sequence of log FB energies have proved to be competitive with respect to mel-cepstrum [6] [7]. However, frequency filtering not only can be performed on log FB energies, but it can also be applied when an LP analysis is performed, as it is described in [6], even in the case of the hybrid spectral analysis described in section 2. Finally, it is worth noting the computational simplicity of filtering with respect to the mel-cepstrum representation. A way of further reducing the computations is to use the filters $1-z^{-1}$ and $z-z^{-1}$, since they do not need products and avoid the average subtraction due to their zero at zero quefrequency. The first filter $1-z^{-1}$, that is equivalent to a slope lifter [5], just consists in subtracting the last log band energy to the current one. The second filter $z-z^{-1}$, that is equivalent to the so-called band-pass liftering [3], consists in subtracting the two log band energies adjacent to the current one. Those simple filters do not depend on the database and they seem to yield speech recognition results close to those of the optimal filter, which is data base dependent [6]

5. IDENTIFICATION EXPERIMENTS

We carried out text-independent speaker identification experiments in both clean and noisy conditions by filtering the average-subtracted frequency sequence of log FB energies in several ways, and using the filtered sequence as speech representation, with no addition of supplementary differential features.

The TIMIT database was used in our experiments. 200 speakers (100 male and 100 female) were selected. Clean speech was used for training in all the experiments. Noisy speech for testing was simulated by adding zero mean white Gaussian noise to the clean signal so that the SNR of the resulting signal becomes 20 dB.

The HTK software, based on the Continuous-Density Hidden Markov Models (CDHMM), was modified to perform speaker recognition experiments with the novel speech representation. In the parameterization stage, after pre-emphasizing the signals with a zero at $z=0.95$, Hamming windowed frames of 25 ms were taken every 10 ms. Each frame was represented by $M=20$ parameters, derived from a bank of $Q=20$ filters. Each speaker was characterized by a Markov model of one state with 32 mixtures with diagonal covariance matrices. The silence was also characterized by a Markov model, but with 3 states and only one mixture. For each speaker, the model was trained with 5 TIMIT sentences. The other 5 TIMIT sentences were used separately as test signals.

Table 1 shows the speaker identification rates (ID) in clean and noisy conditions obtained with the conventional mel-cepstrum coefficients (MFCC) representation along with the ones obtained with the filtered log FB energies (FLFBE) using several high-pass first order FIR filters: $1-0.75z^{-1}$, which equalizes the TIMIT database for $M=Q=20$, as it is used in the previous figures; and $1-0.8z^{-1}$, $1-0.9z^{-1}$ and $1-z^{-1}$, that are inspired by the increasing tilt of the curve of the ratio between inter-speaker and global cepstrum variances in Figure 2.

Parameters / ID	clean	20 dB
MFCC	98.1	32.4
FLFBE ($1-0.75z^{-1}$)	98.3	46.1
FLFBE ($1-0.8z^{-1}$)	98.5	52.8
FLFBE ($1-0.9z^{-1}$)	98.4	61.8
FLFBE ($1-z^{-1}$)	98.3	64.4

Table 1. Speaker identification rates

It can be seen in Table 1 that the new FLFBE parameterization is competitive with conventional MFCC representation in clean conditions. When the optimal equalizer for the TIMIT database $1-0.75z^{-1}$ is used, FLFBE outperforms conventional mel-cepstrum. However, the best results are obtained by using the filter $1-0.8z^{-1}$, which slightly overemphasizes higher quefrecencies with respect to the equalized cepstrum.

The simple database-independent filter z^{-1} , that yielded results close to those of the optimum filter in clean speech recognition [6], has not provided so good results in this case, 97.8 % identification rate. It can be due to the band-pass characteristics of this filter. Actually, high-pass filters, like the ones considered in the Table 1, are more convenient in order to emphasize properly the higher quefrecencies.

Regarding to noisy conditions, excellent results have been obtained by using the new FLFB approach. Using the optimum equalizer $1-0.75z^{-1}$, there is an identification error rate reduction of almost 30 % respect to conventional mel-cepstrum. The results are even better by using filters that put more emphasis on higher quefrecencies. Setting the zero at $z=1$ (filter $1-z^{-1}$), there is an identification error rate reduction of almost 50 %. Filters with zero close to 1 are more convenient in the presence of broad-band noise due to the fact that cepstral parameters of lower index are globally more affected by this type of noise than higher order ones.

6. VERIFICATION EXPERIMENTS

In order to complement those experiments, text-dependent speaker verification experiments on the new speaker-oriented telephone-line POLYCOST database have been performed by combining frequency filtering and hybrid spectral analysis.

The POLYCOST is a new speaker-oriented database that has been recorded as a common initiative within the European COST 250 action entitled 'Speaker Recognition in Telephony'. The database was collected through the European telephone network during January-March 1996. The recording has been performed with an 8 kHz sampling rate.

It contains around 10 sessions recorded by 134 subjects from 14 countries. The majority of non native English speakers gives the possibility to experiment intra-, inter-speaker, language and country variabilities. One session is set up of 15 prompts including one prompt for DTMF detection, 10 prompts with connected digits uttered in English, 2 prompts with sentences uttered in English and 2 prompts in mother tongue.

A set of baseline experiments has been defined. The text-dependent speaker verification experiment has been chosen in this work. The task in this experiment is speaker verification on a fixed password phrase, which is common to all speakers, concretely, "Joe took father's green shoe bench out".

A client model per speaker has been built from the first 4 sessions. A world-model has been built from the first 5 sessions of 22 speakers that have been set aside as an off-line database.

True-identity tests are made on 5th session and later sessions. With existing sessions for 110 client speakers

chosen, this gives 666 true-identity tests. To simulate impostor attempts against speaker X, the 5th session from all speakers in the database except speaker X is used. With 109 impostor tests per client, there are 11990 impostor tests in the experiment.

For comparison purposes, a common scoring software is used, which was developed within the CAVE project. This implementation is based on the methods described in the EAGLES handbook [12].

The HTK recognition system was also used as a recognizer with the same model topology. In the parameterization stage, the speech signal (non-preemphasized) was divided into frames of 20 ms at a rate of 10 ms, and each frame was characterized by $M=20$ parameters obtained by any of the spectral analysis techniques considered above -LP, FB, LP-FB, FB-LP-, and using either cepstral transformation or frequency filtering. When an LP analysis was performed, the prediction order was fixed to 20. Also when a FB was used the number of filters was fixed to 20. Only static parameters were used, neither energy nor delta-parameters.

Table 2 shows the speaker verification results in terms of equal error rate (ERR) obtained with the conventional spectral analysis FB and LP techniques and also the two hybrid approaches, FB-LP and LP-FB, both using cepstral transformation and frequency filtering of log band energies.

Several high-pass first order FIR filters have been considered: $1-0.5z^{-1}$, which equalizes the variance of mel-cepstrum in the isolated digit utterances of the adult portion of the TI database for $M=Q=12$, as it is used in [6]; $1-0.75z^{-1}$, which equalizes the variance of mel-cepstrum in the TIMIT database for $M=Q=20$; and $1-z^{-1}$, that is inspired by the increasing tilt of the curve of Figure 3. Also the band-pass second order filter $z-z^{-1}$ has been tested.

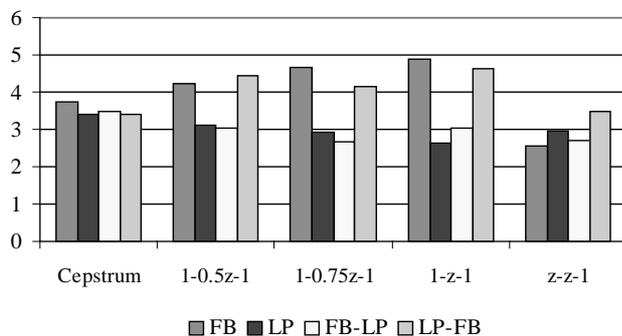
The speaker verification results in terms of EER of Table 2 have also represented graphically in Figure 4. As it can be seen, when the conventional cepstral transformation is used, the best results are obtained by using linear prediction analysis. Standard LP-cepstrum obtains a 3.396 % EER, meanwhile conventional mel-cepstrum gives a 3.748 % EER. Hybrid spectral analysis yield intermediate results: FB-LP-cepstrum gives 3.476 % EER and LP-FB cepstrum gives 3.405 % EER.

Analysis	Cepstrum	$1-0.5z^{-1}$	$1-0.75z^{-1}$	$1-z^{-1}$	$z-z^{-1}$
FB	3.748	4.223	4.674	4.883	2.546
LP	3.396	3.095	2.921	2.648	2.961
FB-LP	3.476	3.044	2.684	3.045	2.706
LP-FB	3.405	4.444	4.142	4.629	3.463

Table 2. Speaker verification results in terms of EER

% EER

Figure 4. Graphic representation of speaker verification results.



Regarding to the use of the new frequency filtering technique, the results depend drastically on the type of spectral analysis. In the case of FB analysis, the only filter that outperforms cepstral transformation is the second-order band-pass filter $z-z^{-1}$. This result do not agree with the previous conclusions about the convenience of high-pass filters. However, using this band-pass filter a 2.546 % EER is achieved, the best results among all those obtained in this work. The relative improvement with respect mel-cepstrum is about 32 %.

In the case of LP spectral analysis, all of the first-order high-pass filters outperform cepstral transformation. The best result, a 2.648 % EER, is obtained by using $1-z^{-1}$. In this case, the relative improvement with respect LP-cepstrum is about 22 %.

With respect to the hybrid spectral analysis, the performance of frequency filtering is quite different. In the case of LP-FB analysis, the use of frequency filtering does not outperform cepstral transformation. Regarding to the FB-LP analysis, a 2.706 % EER is obtained by using the band-pass filter $1-z^{-1}$ and a 2.684 % EER by using the high-pass filter $1-0.75z^{-1}$. In this case, the relative improvement with respect FB-LP-cepstrum is about 23 %.

7. CONCLUSIONS

In this paper, two ways of obtaining more robust parameters have been explored for speaker recognition: the hybridization of both linear prediction (LP) and filter-bank (FB) analysis, and the frequency filtering of log band energies as an alternative to cepstrum. This combination, that had shown to be able to outperform conventional techniques in clean and noisy word recognition, has yield good speaker identification and verification scores. The best verification results on the new speaker-oriented telephone-line POLYCOST database have been obtained by using a second-order

band-pass filter for FB spectral analysis and a first-order high-pass filter for LP and FB-LP (FB prior to LP) analysis.

ACKNOWLEDGMENTS

The author want to acknowledge Climent Nadeu for his suggestions, and to Jordi Muñoz and Javier Martín for their help in the software development.

REFERENCES

- [1] J. Thompson, J.S. Mason, "Within Class Optimization of Cepstra for Speaker Recognition", Proc. EUROSPEECH'95, pp. 165-168.
- [2] J. Hernando, C. Nadeu, "Robust Speech Parameters Located in the Frequency Domain", Proc. EUROSPEECH'97, pp. 417-420.
- [3] B.H. Juang, L.R. Rabiner, J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition", Proc. ICASSP'86, pp. 765-8.
- [4] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition", Proc. ICASSP'86, pp. 761-4.
- [5] B.A. Hanson, H. Wakita, "Spectral Slope Based Distorsion Measures for All Pole Models of Speech", Proc. ICASSP'86, pp. 757-60.
- [6] C. Nadeu, J. Hernando, M. Gorricho, "On the Decorrelation of Filter-Bank Energies in Speech Recognition", Proc. EUROSPEECH'95, pp. 1381-1384.
- [7] J. Hernando, C. Nadeu, "CDHMM Speaker Recognition by means of Frequency Filtering of Filter-Bank Energies", Proc. EUROSPEECH'97, pp. 2363-2366.
- [8] J. Hernando, C. Nadeu, "Speaker Verification on the POLYCOST database using frequency filtered spectral energies, Proc. ICSLP'98, pp. 129-132.
- [9] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", JASA, Vol. 87, No. 4, pp. 1738-52, 1990.
- [10] M.G. Rahim, B.H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition", IEEE Trans. SAP, Vol. 4, No. 1, pp. 19-30, 1996.
- [11] J.W. Picone, "Signal Modeling Techniques in Speech Recognition", Proc. IEEE, Vol.81, No.9, Sept.1993, pp. 1215-47
- [12] F. Bimbot, G. Chollet, "Assesment of Speaker Verification Systems", In.: *Spoken Resources and Assessment, EAGLES Handbook*