A UNIFIED PARAMETERIZATION SCHEME FOR NOISY SPEECH RECOGNITION

Javier Hernando, Climent Nadeu

Universitat Politècnica de Catalunya
Barcelona, Spain
javier@gps.tsc.upc.es

ABSTRACT*

LP-based and mel cepstrum coefficients are by far the most prevalent parameterization techniques in speech recognition. The conventional LP technique is known to be very sensitive to the presence of additive noise and there are few comparative studies about its relative robustness to noise with respect to mel-cepstrum. In this paper, a new unified parameterization scheme is proposed that combines both LP and mel-scaled filter bank analysis and yields new hybrid methods. Also it has been considered the frequency filtering of log filter-bank energies proposed recently by the authors as an alternative to cepstrum representations. From the comparison between conventional and new techniques, excellent results are presented in CDHMM clean and noisy digit recognition.

1. INTRODUCTION

In speech recognition, the short-time spectral envelope of every speech frame is often represented by a set of cepstral coefficients *C(m)*, which are the Fourier series coefficients of its logarithm. These cepstral coefficients usually come either from a set of mel-scaled log filter-bank (FB) energies *–mel-cepstrum–* or from a linear prediction (LP) analysis *–LP-cepstrum* [1].

The conventional LP technique is known to be very sensitive to the presence of additive noise. This fact yields poor recognition rates in noisy conditions when LP-cepstrum are used. Unfortunately, there are few comparative studies about the relative robustness to noise of mel-cepstrum with respect to LP-cepstrum. Actually, one of the main attempts to combat the noise problem consists of finding novel acoustic representations that are resistant to noise corruption in order to replace the traditional parameterization techniques [2].

On the other hand, the authors have recently shown [3] that a discriminative quefrency weighting can be achieved by somewhat decorrelating the frequency sequence of log FB energies with a computationally inexpensive filter, and that the spectral parameters that result from this kind of frequency filtering are competitive with respect to the conventional cepstrum.

* This work has been supported by the grants TIC 95-1022-C05-03 and TIC 95-0884-C04-02

The aim of this paper is threefold: 1) to present a new unified parameterization scheme, that combines LP and mel-scaled FB spectral estimation and includes those conventional techniques as particular cases; 2) to compare the performance of the conventional LPC and mel-cepstrum representations and two new hybrid methods, that are also particular cases of that unified scheme, in CDHMM clean and noisy digit recognition; and 3) to gain some perspective of the merit of the frequency filtering of log FB energies approach in this task.

2. A UNIFIED PARAMETERIZATION SCHEME

Linear prediction, which is equivalent to an AR spectral modeling, is widely used in speech processing and, particularly, in speech recognition. Concretely, it has been shown that the use of the liftered LP-cepstrum (LP-C) in the conventional Euclidean distance measure leads to the best results of those obtained with that model. The strength of this 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 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 traditionally employed- 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 has several advantages derived from the frequency localization of the parameters. Concretely, it allows straightforward techniques that cope with noise. Mel-cepstrum, probably the most used parameters in speech recognition [1], come from this FB approach. Hereafter, we will refer to mel-cepstral coefficients as FB-C (filter-bank cepstrum).

The combination of LP and FB analysis may yield improved cepstral parameters. One possible approach is to apply FB analysis on the signal prior to LP analysis [4] [5]. In this work, the corresponding cepstrum will be referred to as FB-LP-C and it is computed similarly to the PLP coefficients [4], but using a higher order LP

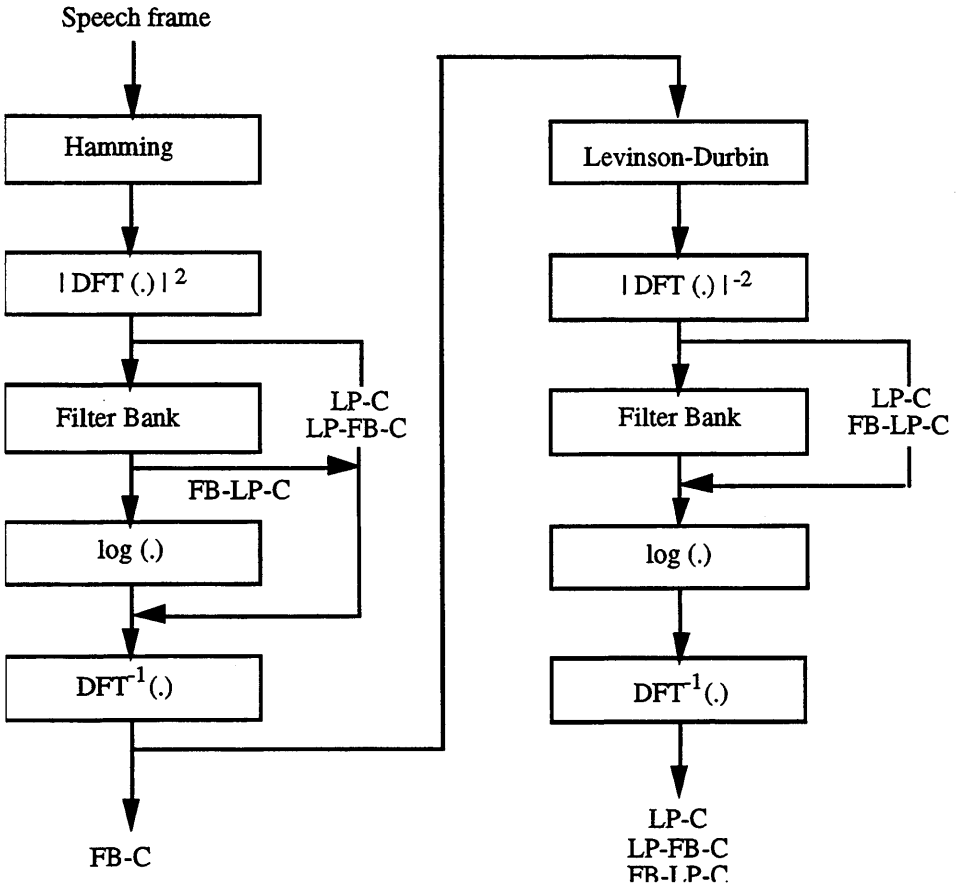

Fig.1. Block-diagram of the unified parameterization scheme

analysis without perceptual weighting and amplitude compression. Another approach proposed in this work is to apply LP analysis followed by FB analysis to give what will be referred to as LP-FB-C.

Both conventional LP-C and FB-C representations and the two new hybrid FB-LP-C and LP-FB-C methods can be considered as particular cases of the unified parameterization scheme of the Figure 1. Furthermore, combining LP and FB spectral estimation, this scheme can lead to novel speech parameterization techniques.

3. FREQUENCY FILTERING OF FILTER-BANK ENERGIES

The sequence of cepstral coefficients $C(m)$ is a quasi-uncorrelated and compact representation of speech spectra. Actually, the quefrency sequence is always windowed [1] to eliminate the cepstral coefficients beyond a quefrency M. And, for some type of speech recognition systems, the window also appropriately weights the remaining coefficients. In this case, two steps are needed for obtaining the final parameters from the log FB energies or the LP coefficients: 1) a linear transformation, that significantly decorrelates the sequence of parameters, and 2) a discriminative weighting of the cepstral coefficients. Additionally, in continuous observation Gaussian density HMM (CDHMM) with diagonal covariance matrices, the shape of the cepstral window has no effect due to the intrinsic variance normalization of the exponent of the Gaussian pdf. So only its length, i.e. the number of parameters M, is a control variable.

In recent papers [3] [6], in order to try to overcome those disadvantages, the authors have presented an alternative to the use of cepstrum in speech recognition that consists of a simple linear processing on the log FB 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 (FF) to denote that the convolution is performed on the frequency domain. As shown in [6], FF produces both effects, decorrelation and discrimination, in only one step and using an extremely simple first or second order FIR filter. Besides, FF is able to produce an implicit cepstral weighting in CDHMM with diagonal covariance matrices.

Hereafter, we will refer to frequency filtered log FB energies as FB-F to distinguish from FB-C (mel-cepstrum). In this work FF is also applied when an LP analysis is performed, as it is described in [3], to give what we will call LP-F. In the same manner, we will consider the hybrid methods proposed in section 2 to obtain what we will call BF-LP-F and LP-BF-F.

4. RECOGNITION EXPERIMENTS

4.1. Database and Recognition System

The database used in the recognition experiments consists of 20 repetitions of the English digits corresponding to the adult speakers (112 for training and 113 for testing) of the speaker independent digit TI [7] database. The initial sampling frequency 20 kHz was converted to 8 kHz. 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, 10 and 0 dB.

The HTK recognition system, based on CDHMM, was appropriately modified and used for the recognition experiments. In the parameterization stage, the speech signal (non-preemphasized) was divided into frames of 30 ms at a rate of 10 ms, and each frame was characterized by M parameters obtained by any of the parameterization techniques described above. Only static parameters were used, neither energy nor delta-parameters. Each digit was characterized by a first order, left-to-right, Markov model of 10 states with one mixture of diagonal covariance matrix and without skips. The same structure was used for the silence model but only with 5 states. Training was performed in two stages using Segmental k-means, with previous manual endpointing, and Baum-Welch algorithms.

4.2. Experimental Results

Figure 2 shows the digit recognition rates obtained for all parameterization techniques described in this paper in clean conditions and for 20, 10 and 0 dB of additive white noise. The results corresponding to cepstrum representations are in the left column, and the results obtained by frequency filtering of log energies are in the right one. For each parameterization technique and level of noise, the number of parameters M was varied from 8 to 20. When an LP analysis was performed, the prediction order was always fixed to M. Regarding to the number of the filters of the FB, it was fixed to 20 except for the FB-F and LP-FB-F front-ends, in which the number of filters is equal to M.

In clean conditions, it can bee seen in Figure 2.a that conventional mel-cepstrum (FB-C) outperforms clearly LP-C. FB-C obtains good results with low values of M and LP-C only yields good recognition rates with M values higher than those commonly used in speech recognition. The best recognition rates are obtained by FB-C with low values of M. Regarding to the hybrid methods, LP-FB-C obtains intermediate results between conventional techniques and it is not sensitive to the value of M. On the other hand, the FB-LP-C representation outperforms clearly conventional techniques.

As it can be seen in figure 2.e, also in clean conditions, FF yields excellent results for FB-F and FB-LP-F representations, especially for intermediate values of M. For LP-F and LP-BF-F, only good results are obtained for high values of M. For these tests, we used the most simple filter proposed in [3], i.e. $H(z)=z-z^{-1}$.

Regarding to the robustness of the various techniques to additive white noise, very good results are also obtained. As it can be seen in the figures 2.b-d, FB-LP-C yields the best results among cepstrum representations followed by conventional FB-C. The difference between both methods decreases for severe noisy conditions. On the other hand, figures 2.f-h show that FF approach yields excellent results for FB-LP-F at moderate levels of noise, and for FB-C in severe noise conditions.

5. CONCLUSIONS

Combining both LP and filter-bank analysis, we have proposed a new unified parameterization scheme that includes conventional LP and mel-cepstral representations and yields new hybrid methods. In our experiments, conventional mel-cepstrum outperforms LP-cepstrum in both clean and noisy conditions. Even better results have been obtained by applying FB analysis prior to LP analysis also in both clean and noisy conditions. A few initial experiments using noisy signals obtained from a real application seem to indicate the application of LP analysis followed by FB analysis outperforms the other representations. Finally, the best results have been obtained by using the frequency filtering approach, recently proposed by the authors [3] as an alternative to cepstrum.

ACKNOWLEDGMENTS

The authors would like to thank Pascual Ejarque for his help in software development.

REFERENCES

[1] J. W. Picone, "Signal modeling techniques in speech recognition", Proc. IEEE, Vol. 81, No. 9, pp. 1215-47, 1993.

[2] B.H. Juang, "Speech recognition in adverse environments", Computer Speech and Language", Vol. 5, pp. 275-294, 1991.

[3] C. Nadeu, J. Hernando, M. Gorricho, "On the decorrelation of fiter-bank energies in speech recognition", Proc. EUROSPEECH'95, pp. 1381-1384.

[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", JASA, Vol. 87, No. 4, , pp. 1738-1752, 1990.

[5] M.G. Rahim, B.H. Juang, "Signal bias removal by maximum linelihood estimation for robust telephone speech recognition", IEEE Trans. SAP, Vol. 4, No. 1, pp. 19-30, 1996.

[6] C. Nadeu, J.B. Mariño, J. Hernando, A. Nogueiras, "Frequency and time-filtering of filter-bank energies for HMM speech recognition", Proc. ICSLP'96, pp. 430-436.

[7] R.G. Leonard, "A database for speaker-independent digit recognition", Proc.ICASSP'84, pp. 42.11.1-4.

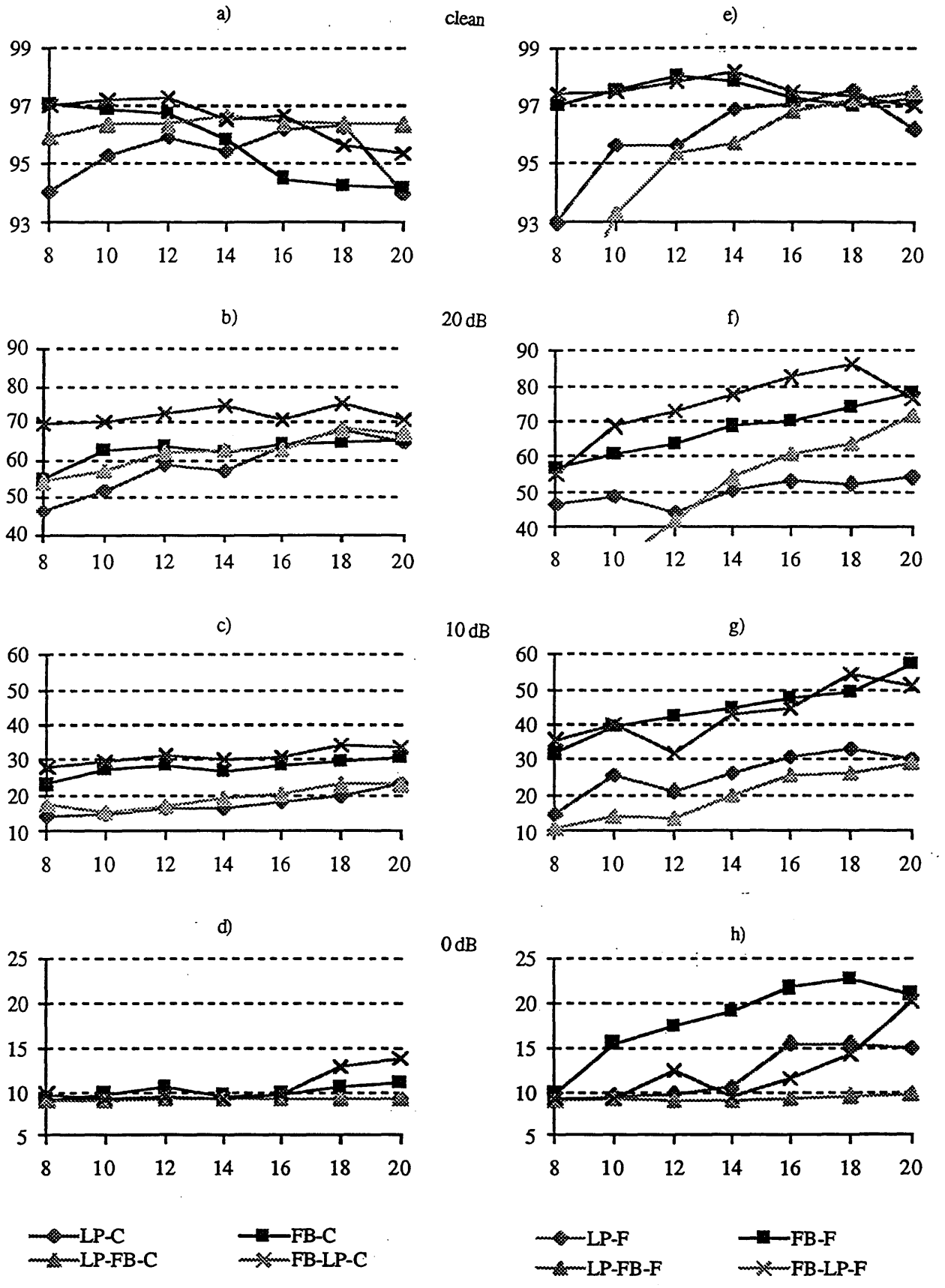

Figure 2. Digit recognition rates, varying the number M of parameters, for the various cepstrum representations (left column) and the frequency filtering approach (right column) in clean conditions, 20, 10 and 0 dB of additive white noise.