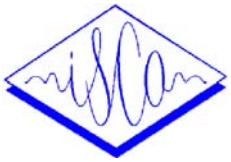

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

ROBUST SPEECH PARAMETERS LOCATED IN THE FREQUENCY DOMAIN

J. Hernando and C. Nadeu
Universitat Politècnica de Catalunya
Barcelona, Spain
javier@gps.tsc.upc.es

ABSTRACT*

In this paper, two ways of obtaining more robust spectral parameters are explored. Firstly, an hybridization of both LP and filter-bank approaches is considered, which is capable of improving recognition results for both noisy and clean speech in CDHMM digit recognition. Secondly, better performance may also be achieved by replacing the cepstral coefficients by a recently proposed set of parameters located in the frequency domain which come from a simple filtering of the log band energies.

1. INTRODUCTION

In speech recognition, the short-time spectral envelope of every speech frame is usually represented by a set of cepstral coefficients C(m), $1 \le m \le M$, which are the Fourier series coefficients of its logarithm. These 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. So it yields poor recognition rates in noisy conditions when LP-cepstrum is used. The authors have considered in the past the one-sided autocorrelation LP (OSA-LP) representation, based on the LP in the autocorrelation domain [3].

Unfortunately, there are few comparative studies about the relative robustness to noise of mel-cepstrum with respect to LP-cepstrum. Recently, the authors have considered a unified parameterization scheme that combines both LP and filter-bank analysis [4].

On the other hand, the authors have recently shown [5] [6] that the set of parameters located in the frequency domain that result from filtering the frequency sequence of band energies with a simple FIR filter of order 1 or 2 is competitive with respect to the conventional cepstrum coefficients for clean speech.

The aim of this paper is to gain some perspective of the merit of all those techniques in both clean and noisy speech recognition. In sections 2 and 3, the unified parameterization scheme and the frequency filtering technique will be briefly revised. Section 4 is dedicated to show the experimental results obtained by applying these techniques in CDHMM clean and noisy isolated digit recognition, with both white noise and noise from a real task.

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

2. A UNIFIED PARAMETERIZATION SCHEME

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 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 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 masking). Mel-cepstrum, probably the most used parameters in speech recognition [1], come from this FB approach.

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 [7] [8]. It will be referred to as FB-LP and it is computed similarly to the PLP coefficients [7], 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 cepstrum (LP-C) and mel cepstrum (FB-C) parameterizations and the cepstrum representations corresponding to the two new hybrid FB-LP and LP-FB methods (FB-LP-C and LP-FB-C, respectively) are encompassed in the unified parameterization scheme of the Figure 1, where Filter Bank refers to the band integration stage. Furthermore, combining LP and FB spectral estimation, this scheme can lead to other novel speech parameterization techniques.

3. FREQUENCY FILTERING OF BAND ENERGIES

The sequence of cepstral coefficients C(m) is a quasi-uncorrelated and compact representation of speech spectra. Actually, the quefrency sequence is always

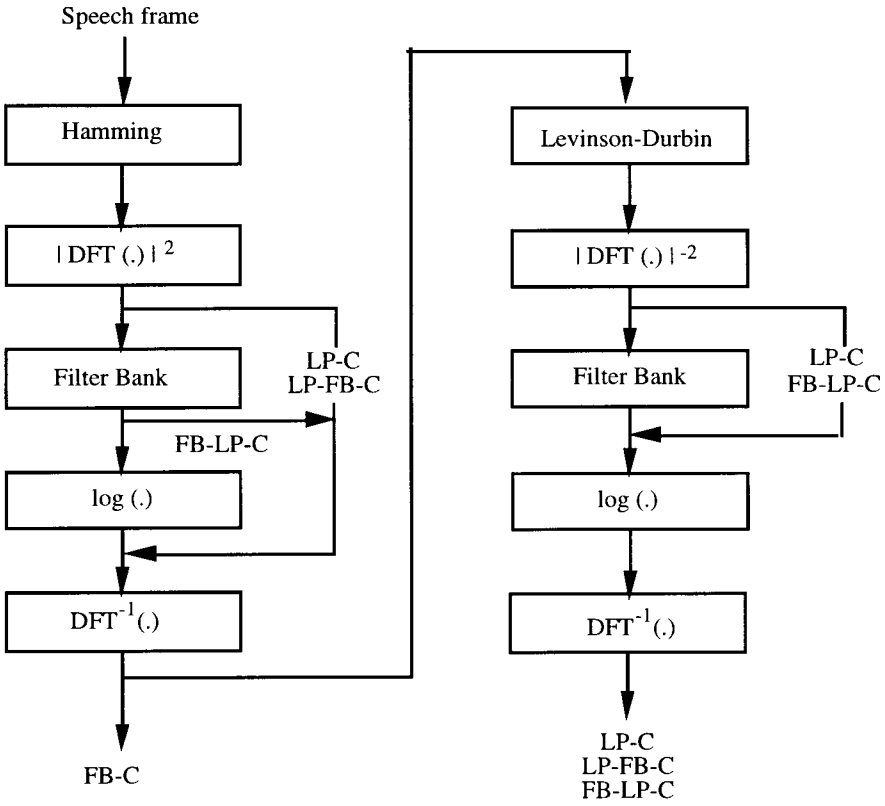

Fig.1. Block-diagram of the unified parameterization scheme for cepstral representations

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 (liftering). In the latter 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 variance normalization included in the exponent of the Gaussian pdf. So only the window length is a control variable.

In recent papers [5] [6], in order to try to overcome those disadvantages and to have parameters that possess a frequency meaning, an alternative to cepstral coefficients has recently been introduced. It consists in a simple linear processing in the log band energy domain. The transformation of the sequence of log band 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. FF not only can be performed on FB energies, but it can also be applied when an LP analysis is performed, as it is described in [5]. 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.

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 [9] 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 and also low-pass noise from a real task to the clean signal.

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 analysis techniques considered above, LP, FB, LP-FB, FB-LP, OSA-LP, and also an hybrid OSA-LP-FB, and using either cepstrum transformation or frequency filtering. Cepstrum representations will be referred to as with the suffix -C and frequency filtering representations will be denoted using the suffix -F. 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, LP-FB-F and OSA-LP-FB-F front-ends, in which the number of filters is equal to M.

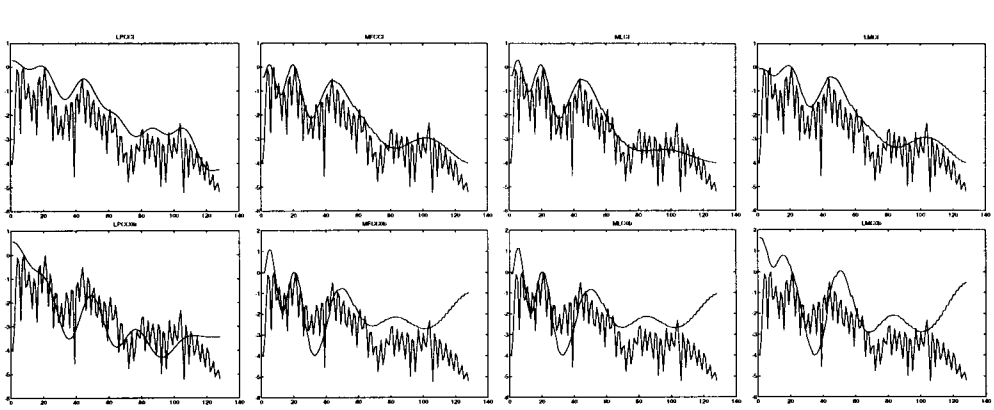

Fig. 2. Spectrum estimates of a voiced speech frame in noise free conditions (top row) and 0 dB SNR of additive white noise (bottom row) for LP, FB, FB-LP and LP-FB analysis techniques. The clean periodogram estimate is also drawn in all graphs for comparison purposes.

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 spectrum estimates of a voiced speech frame in noise free conditions (top row) and 0 dB SNR of additive white noise (bottom row) for LP, FB, FB-LP and LP-FB analysis techniques. These spectrum estimates have been computed as the DFT of the zero-padded corresponding cepstrum. An inverse mel transformation has been applied to compare with the clean periodogram when FB analysis has been performed. It can be seen in the figure that FB-LP estimates are very similar to FB ones, whereas LP-FB estimates are in between LP and FB ones.

Figure 3 shows the digit recognition rates obtained in clean conditions for even values of M from 8 to 20. Two frequency filters have been considered: $z\text{-}z^{-1}$, the same used in [5], that is equivalent to a band-pass liftering; and $1\text{-}z^{-1}$, that is similar to a slope lifter. It can be seen in Figure 3.a that conventional mel-cepstrum (FB-C) outperforms clearly conventional LP-C with lower values of M. Regarding to the hybrid methods, LP-FB-C obtains intermediate results between both conventional techniques and it is not sensitive to the value of M. However, FB-LP-C outperforms clearly both conventional techniques. Figure 3.b shows that FF by using the filter $z\text{-}z^{-1}$ yields good results for FB-F and FB-LP-F representations, but the use of the filter $1\text{-}z^{-1}$ does not achieve any clear improvement with respect to cepstrum representations. The rates of OSA-LP-based techniques do not appear in Figure 3 since they lie just under 93 %.

Figure 4 shows the results for 20, 10 and 0 dB of additive white noise by using cepstrum representations and FF with the filter $1\text{-}z^{-1}$. As it can be seen in the figures 4.a-c, OSA-LP-based techniques yield the best results among cepstrum representations. FB-LP-C provides the best results among the other cepstrum representations followed by conventional FB-C. On the other hand, Figures 4.d-f show that the use of the filter $1\text{-}z^{-1}$ yields better results than the corresponding cepstral representations, especially for FB-LP-F at moderate levels of noise, but not for OSA-LP-based techniques. The results obtained by this high-pass filter are better than those obtained by using the band-pass filter $z\text{-}z^{-1}$ used in [4]. It is due to the fact that cepstral parameters of lower index are more affected by this type of noise than higher order ones.

Finally, Figure 5 shows the results for 20, 10 and 0 dB of real low-pass noise for cepstrum representations. As it can be seen, the best results are obtained in this case by using the LP-FB-C representation at moderate SNR, and OSA-LP-FB-C at low SNR. The frequency filters considered above have not outperformed cepstrum representations for this type of noise.

5. CONCLUSION

The alternative parameterizations considered in this work have shown to be able to outperform the conventional LP and mel cepstrum for both clean and noisy speech in CDHMM isolated word recognition. The application of filter-bank analysis prior to LP analysis outperforms both conventional approaches for both clean speech and additive white noise, whereas the application of LP analysis followed by filter-bank analysis is preferable for the real low-pass noise used in this work. The application of LP in the autocorrelation domain may be a proper choice in noisy conditions. Finally, frequency filtering of log band energies as an alternative to cepstrum lead to good results by using the band-pass filter $z\text{-}z^{-1}$ for clean speech, and the high-pass filter $1\text{-}z^{-1}$ for additive white noise. Further work is needed to design a suitable filter operation for real noises.

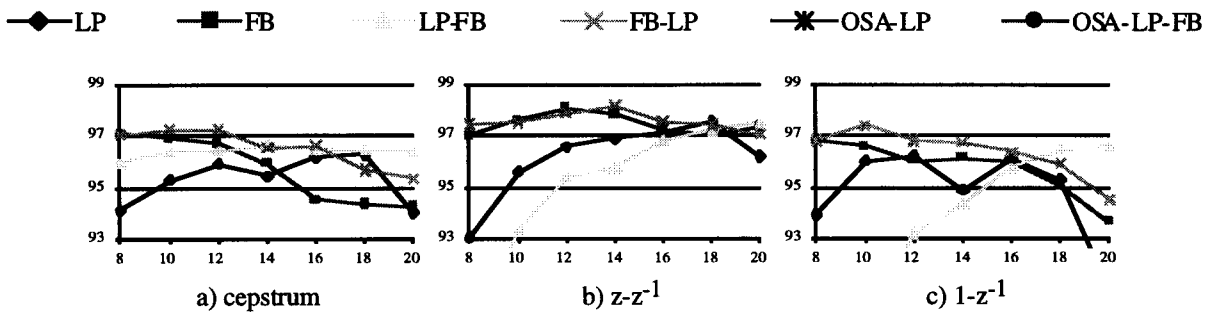

Figure 3. Digit recognition rates in clean conditions.

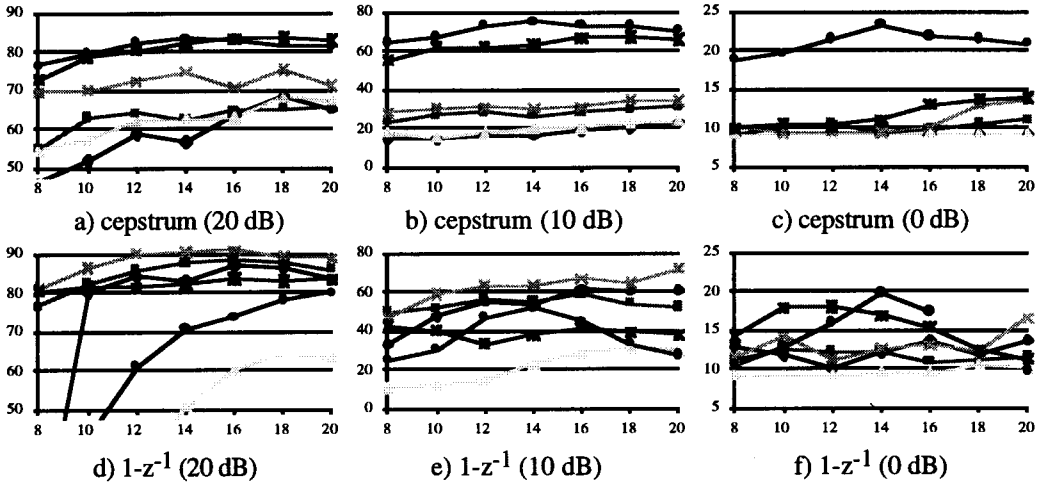

Figure 4. Digit recognition rates adding white noise.

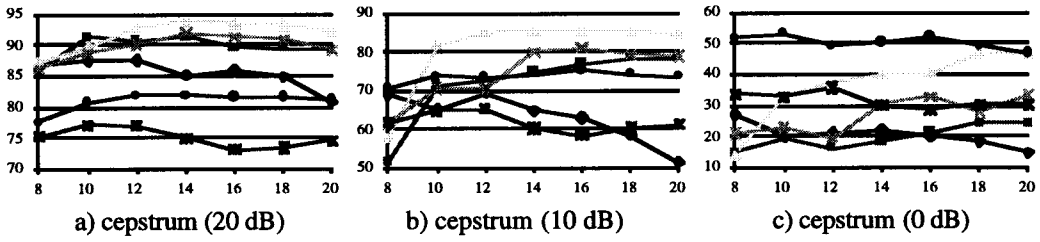

Figure 5. Digit recognition rates adding real low-pass noise.

6. ACKNOWLEDGMENTS

The authors would like to thank P. Ejarque and D. Company for their help in software development

7. 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. 275-94, 1991.

[3] J. Hernando, C. Nadeu, "Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques", Proc. ICASSP'94, pp. II-69-72.

[4] J. Hernando, C. Nadeu, "A unififed parameterization scheme for noisy speech recognition", Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pony-à-Mousson, France, April 1997, pp. 115-8.

[5] C. Nadeu, J. Hernando, M. Gorricho, "On the decorrelation of fiter-bank energies in speech recognition", Proc. EUROSPEECH'95, pp. 1381-4.

[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. 430-6.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", JASA, Vol. 87, No. 4, , pp. 1738-52, 1990.

[8] 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.

[9] R.G.Leonard, "A database for speaker-independent digit recognition", Proc.ICASSP'84, pp. 42.11.