ON THE USE OF FILTER-BANK ENERGIES DRIVEN FROM THE AUTOCORRELATION SEQUENCE FOR NOISY SPEECH RECOGNITION

Javier Hernando
TALP Research Center, UPC, Barcelona, Spain
javier@talp.upc.es

ABSTRACT*

The OSA-LP (One-Sided Autocorrelation LP) representation of speech signal has shown to be attractive for noisy speech recognition because of both its high recognition performance with respect to the conventional LP in severe conditions of additive broadband noise and its computational simplicity. However, the mel-cepstrum representation, which comes from a filter bank (FB) analysis, has become increasingly popular. Furthermore, hybrid techniques that combine filter-bank and LP analysis, have also been proposed. The aim of this paper is to gain some perspective of the merit of the autocorrelation, LP and FB based techniques in a real environment.

1. INTRODUCTION

In speech recognition, the short-time spectral envelope of every speech frame is often represented by a set of cepstral coefficients $C(m)$, that 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]. 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].

The conventional LP technique is known to be very sensitive to the presence of additive noise. This fact leads to poor recognition rates in noisy conditions when this technique is used, even if only a moderate level of contamination is present in the speech signal. As an alternative representation of speech signals when noise is present, the authors have considered in the past the One-Sided Autocorrelation LP (OSA-LP) representation. This technique is essentially an AR modeling of the causal part of the autocorrelation sequence, and it has shown to be attractive in noisy speech recognition because of both its computational simplicity and its high recognition performance with respect to standard LP in severe conditions of additive white noise [3] and broadband noises [4].

However, mel-cepstrum representation, which comes from a filter bank (FB) analysis, has become increasingly popular due perhaps to the higher flexibility of the sub-band approach. Furthermore, hybrid techniques that combine FB and LP approaches have also been proposed [5] [6]. Recently, the authors have considered a unified parameterization scheme for noisy speech recognition [7] [8], which combines both LP and FB analysis.

The aim of this paper is to gain some perspective of the performance of the autocorrelation, LP and FB based techniques in a real environment. In sections 2 and 3 the OSA-LP representation and the hybrid techniques will be revised. Section 4 is devoted to show the experimental results obtained by applying these techniques on the TI and Aurora 1.0 databases. The later is being used for developing the ETSI noisy Distributed Speech Recognition (DSR) standard front-end [9].

2. OSA-LP REPRESENTATION

From the autocorrelation sequence $R(n)$, we may define the one-sided autocorrelation sequence $R^+(n)$ as its causal part

$$R^+(n) = \begin{cases} R(n) & n > 0 \\ R(0)/2 & n = 0 \\ 0 & n < 0 \end{cases}$$ (1)

Its Fourier transform is the complex spectrum

$$S^*(\omega) = \frac{1}{2} \left[ S(\omega) + j S_H(\omega) \right].$$ (2)

where $S(\omega)$ is the spectrum of the signal, i.e. the Fourier transform of $R(n)$, and $S_H(\omega)$ is the Hilbert transform of $S(\omega)$.

Due to the analogy between $S^*(\omega)$ in (2) and the analytic signal used in amplitude modulation, a spectral "envelope" $E(\omega)$ can be defined as

$$E(\omega) = |S^*(\omega)|$$ (3)

whose square, the square envelope, is the spectrum of $R^+(n)$.

This envelope characteristic, along with the high dynamic range of speech spectra, originates that $E(\omega)$ strongly enhances the

---

*This work has been supported by the grants TIC 98-0685 and TIC 98-0423-C06-01
highest power frequency bands. Thus, the noise components lying outside the enhanced frequency bands are largely attenuated in \( E(\omega) \) with respect to \( S(\omega) \). On the other hand, it is well known that \( R(n) \) has the same poles than the signal itself.

It is then suggested that the AR parameters of the speech signal can be more reliably estimated from \( R(n) \) than directly from the signal itself when it is corrupted by broadband noise. For this purpose, in the same manner as the standard LP performs a linear prediction of the speech signal, that is equivalent to assume an all-pole model for the spectrum of the signal \( S(\omega) \), we may consider a linear prediction of \( R(n) \), equivalent to assume an all-pole model for its spectrum \( E(\omega) \). This is the basis of the OSA-LP (One-Sided Autocorrelation LP) parameterization technique [3][4].

A straightforward algorithm was proposed to calculate the cepstrum coefficients corresponding to the OSA-LP technique, that consists in applying the autocorrelation (windowed) method of linear prediction upon an estimation of the OSA sequence, instead of the signal itself: a) firstly, from the speech frame of length \( N \) the autocorrelation lags until \( M = N/2 \) are estimated (this value of \( M \) was empirically optimized to take into account the well known tradeoff between variance and resolution of the spectral estimate); b) secondly, the Hamming window from \( m = 0 \) to \( M \) is applied; c) thirdly, if \( p \) is the order prediction, the first \( p + 1 \) autocorrelation lags of that OSA sequence are computed from \( m = 0 \) to \( p \) using the conventional biased estimator; d) finally, these values are used as entries to the Levinson-Durbin algorithm to estimate the AR parameters.

3. HYBRID TECHNIQUES

There are many reported techniques to estimate the set of spectral parameters [1], but they always perform some kind of smoothing of raw spectral parameters. Spectral smoothing is used to remove the harmonic structure of the speech spectrum corresponding to pitch information and to reduce the variance (error) of the spectral envelope estimation. That operation has basically been done in two alternative ways: LP analysis and spectral band energy estimation. The strength of LP arises from the fact that it matches the all-pole model speech production. In this way, it is able to approximately separate the vocal response, which corresponds to the spectral envelope, from the glottal excitation.

However, the band energy parameters have become increasingly popular. They separately represent the energy at each band frequency band since they result from integrating the energy values in the time-frequency area specified by the frame length and the effective bandwidth. The main reason of the usefulness of these energies is perhaps the higher flexibility of the sub-band approach with respect to the full-band approach involved in LP modeling. In fact, it offers the possibility of defining the width and shape of the bands along the frequency axis. Also, if the SNR of each band is known, the band energy representation allows using it in straightforward ways: noise masking, spectral subtraction, etc.

The computation of the band energies can be performed in several ways. The classical implementation consists of a bank of filters that perform time convolution followed by a wave rectification and low-pass filtering. Currently, the most used implementation of the filter-bank (FB) analysis operates in the frequency domain by computing the weighted average of the magnitude (or, sometimes, the squared magnitude) of the DFT values of the windowed speech frame in each frequency band, obtaining in this way the so-called filter-bank energies. Mel-cepstrum, probably the most used parameters in speech recognition [1], come from this FB approach.

Hybrid techniques that combine FB and LP analysis have also been proposed [5][6]. The best known one is PLP (Perceptual LP) [5], which applies LP modeling after FB computation and some perceptually motivated processing steps. Note that, as the order of the LP analysis is usually chosen low, the PLP parameterization involves an additional smoothing effect. In this work, FB analysis will be applied on the signal prior to LP analysis, similarly to PLP, but using a higher order LP analysis and without those perceptually motivated steps. It will be referred to as FB-LP. An alternative approach considered in this work is to apply LP analysis followed by FB analysis to give what will be referred to as LP-FB.

Both conventional LP and FB representations and the two new hybrid FB-LP and LP-FB methods can be considered as particular cases of a unified parameterization scheme presented in [7][8]. Furthermore, this scheme can lead to novel speech parameterization techniques.

Logically, the first processing step of the hybrid techniques heavily determines the characteristics of the spectral estimate. In [8] it was observed that the spectral estimates that result from the FB-LP approach are very similar to the FB ones, whereas LP-FB estimates are in between LP and FB ones.

4. EXPERIMENTS AND CONCLUSIONS

4.1. White noise experiments

The database used in the first recognition experiments consists of the isolated digits corresponding to the adult speakers (112 for training and 113 for testing) of the speaker independent digit TI [10] database. The initial sampling frequency 20 kHz was converted to 8 kHz. Clean speech was used for training and 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 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 12 cepstral coefficients obtained by any of the analysis techniques considered above: LP, FB, LP-FB, FB-LP, OSA-LP, and also an hybrid OSA-LP-FB. When LP analysis was performed, the prediction order was always fixed to 12.
When FB analysis was performed the number of filters was fixed to 20. Only static parameters were used, neither energy nor delta-parameters. Each digit was characterized by a 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 end-pointing, and Baum-Welch algorithms.

Figure 1 shows the digit recognition rates for each parameterization technique in terms of the SNR (inf. means clean). The techniques have been ordered according to their average recognition rate: OSA-LP-FB (59.20%), OSA-LP (50.73%), FB-LP (37.76%), FB (34.23%), LP-FB (29.44%) and LP (28.16%). Regarding to the traditional techniques, FB (mel-cepstrum) outperforms clearly LP (LP-cepstrum). However, OSA-LP based techniques yield the best results, particularly the new OSA-based hybrid technique OSA-LP-FB. Note that OSA-based techniques obtain poor results in clean conditions. Regarding to the hybrid methods, LP-FB obtains intermediate results between conventional techniques, but FB-LP outperforms conventional techniques.

Figure 1: Results of white noise experiments.

4.2. Real noises experiments

The database used in the real noises experiments is the Aurora 1.0 database, which is being used for developing the ETSI noisy Distributed Speech Recognition (DSR) standard front-end [9]. This corpus consists of the TI connected digit database downsampled from 20 kHz to 8 kHz with added noises at the following SNR levels: clean, 20, 15, 10, 5, 0 and –5 dB. Four noises have been used: hall, babble, suburban and car. Clean speech was used for training.

The HTK toolkit was also used for the recognition experiments, with only slight differences with respect to the previous experiments: the number of filters for the FB analysis is 23, each digit is modeled by 16 states with 3 mixtures and silence is modeled by 3 states with 6 mixtures.

In the former experiments the best results were obtained by either using the OSA sequence instead of the speech signal itself or applying FB instead/before LP. It is then suggested the application of FB analysis on the OSA sequence, calculated as it was reported in section 2, instead of the signal itself. This front-end, which will be referred to as OSA-FB, did not outperform in our tests the conventional mel-cepstrum (FB). However, the application of FB analysis on the symmetric autocorrelation sequence, from n=-N/2 to N/2, did outperform FB in noisy conditions. This front-end will be referred to as A-FB. Analogously, LP modeling of the symmetric autocorrelation sequence, which will be referred to as A-LP, was tested.

Figure 2 shows the digit recognition rates for these techniques, along with OSA-LP, in terms of the SNR. They have been ordered according to their average recognition rate: A-FB (54.15%), FB (52.07%), OSA-FB (48.17%), OSA-LP (47.08%), LP (27.19%) and A-LP (17.50%). It can be seen clearly that FB-based techniques outperform LP-based ones. It is also clear that OSA preprocessing is only convenient before LP modeling (see justification in section 2), whereas A preprocessing enhances FB. Figures 3 shows the recognition rates in terms of the type of noise. It can be seen that autocorrelation-based preprocessing works better for broad band noises: hall, train and car.

Figure 2: Results for Aurora database of autocorrelation-based parameterizations in terms of SNR.

Figure 3: Results for Aurora database of autocorrelation-based parameterizations in terms of the type of noise.
The autocorrelation sequence can also be computed in the frequency domain as the inverse DFT of the squared magnitude of the DFT of the speech frame. Therefore, if A-FB considered all autocorrelation lags from \( n=-(N-1) \) to \( N-1 \), it would be equivalent to FB with a spectral shaper in the form of a square after the magnitude operation. This shaper can be generalized to a power or other non-linearity and it can be optimized in order to appropriately emphasize/attenuate the spectral peaks/valleys in a noisy environment [11]. Therefore, FB (mel-cepstrum) can be generalized to FB-g, where \( g \) is the exponent value. Figure 4 shows the recognition rates for several values of \( g \) in terms of the SNR. The various front-ends have been ordered according to their average recognition rate in this task, which increases with \( g \) for this range of values, that is, emphasizing the spectral peaks. Figure 5 shows the rates in terms of the type of noise. It is seen that high \( g \) values are only good for broad-band noises.

5. ACKNOWLEDGEMENTS

The author want to acknowledge C. Nadeu and D. Macho for their suggestions, and to P. Eiarque, D. Company and A. Garcia for their help in the software development.

6. REFERENCES