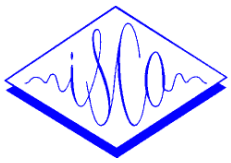


2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

A COMPARATIVE STUDY OF PARAMETERS AND DISTANCES FOR NOISY SPEECH RECOGNITION

Javier Hernando and Climent Nadeu

Dept. of Signal Theory and Communications
Universitat Politècnica de Catalunya
08034 Barcelona, Catalonia, Spain

ABSTRACT

Speech recognition in noisy environments remains an unsolved problem even in the case of isolated word recognition with small vocabularies. Recently, several techniques have been proposed to alleviate this problem. Concretely, the Short-Time Modified Coherence (SMC) parameterization and the Cepstral Projection Distortion (CPD) measure have shown excellent results when tested in a speech recognition system based on Dynamic Time Warping (DTW) and using speech contaminated by additive white noise. In this paper, a new technique based on the AR modeling of the one-sided autocorrelation sequence (OSALPC) is presented and, from a comparative study of these LPC-based techniques in the Hidden Markov Model (HMM) approach, two main conclusions are attained: 1) the slope cepstral window and a relatively high model order are preferable, and 2) the cepstral representation based on the autocorrelation (rather on the signal) modeling achieves excellent results.

1. INTRODUCTION

The performance of existing speech recognition systems degrades rapidly in the presence of background noise when training and testing cannot be done under the same ambient conditions. In order to develop a speech recognition system that operates robustly and reliably in the presence of noise, many techniques have been proposed in the literature for reducing noise in each stage of the recognition process, particularly, in feature extraction and similarity measuring.

A spectral estimation technique widely used in speech processing and, particularly, in speech recognition is linear predictive coding (LPC) [1], equivalent to an AR modeling of the signal. Recent contributions [2] showed that the use of a bandpass liftering (cepstral weighting) of the LPC-cepstral coefficients in the standard Euclidean distance measure can lead to significant improvement in recognition accuracy in noisy free conditions. However, these techniques are not robust to changes of the ambient conditions.

For recognition in noisy speech, Hanson and Wakita [3] used the spectral slope distance measure, which shows high correlation with subjective phonetic distinctions and is equivalent to a quefrency weighting on the cepstral domain. This slope lifter is concerned with the fact that, in the presence of white or broad-band noise, lower order cepstral coefficients are more affected than higher order terms in the truncated cepstral vector.

This work was supported by PRONTIC grant number 105/88

Recently, Mansour and Juang [4] considered that there is no obvious reason to maintain the symmetry characteristics of the Euclidean distance if one knows the reference and test signals have different degrees of noisy corruption. So they proposed a family of robust Cepstral Projection Distortion (CPD) measures based on the vector projection operation which takes into account the effects of additive white noise upon the parameter representation.

The same authors have proposed [5] a new technique for robust spectral analysis of speech called Short-Time Modified Coherence (SMC), based on the well known fact that the autocorrelation sequence is less affected by noise than the original signal. The SMC representation is essentially an AR modeling in the autocorrelation domain and performs better than the traditional LPC for noisy signals in terms of signal-to-noise ratio (SNR) improvement.

The aim of this paper is to make a comparative study of these techniques and to present a parameterization similar to the SMC one, the one-sided autocorrelation linear predictive coding (OSALPC), as a robust representation of speech signals when noise is present.

This paper is organized in the following way. Section 2 is devoted to a brief background about the standard Euclidean distance and the new CPD measures applied upon the liftered cepstral vector. In section 3 the one-sided autocorrelation linear predictive coding (OSALPC) is presented. Section 4 reports the application of these techniques to an isolated word multispeaker recognition task using the HMM approach in order to compare their performance and gain some perspective of the merit of the OSALPC representation in the presence of additive white noise.

2. CEPSTRAL DISTORTION MEASURES

2.1. Euclidean Distance

The Euclidean distance between liftered LPC-cepstral coefficients is defined as follows

$$d_E = \sum_{n=1}^{L} [\, w(n)\, (\, c_t(n) - c_r(n)\,)\,]^2 \qquad (1)$$

where $c_t(n)$ and $c_r(n)$ are the nth cepstral coefficients of the test and reference frames, respectively, L is the number of cepstral coefficients and w(n) is the weight applied to the nth coefficient.

The set of weights w(n) defines a lifter. In this paper we

consider basically two different lifters, reported in [2] and [3] respectively:

$$\text{Bandpass lifter: } w(n) = 1 + \frac{L}{2} \sin\left(\frac{\pi n}{L}\right) \quad (2.a)$$

$$\text{Slope lifter: } w(n) = n \quad (2.b)$$

where n = 1,...,L. If M denotes the LPC model order, the value of L is 3M/2 for the bandpass lifter, and M for the slope lifter.

From liftering, a smoothed version of the spectrum is obtained that depends on both the type of the lifter and the model order. One of the aims in section 4 is to find an optimum degree of smoothing in noisy conditions.

2.2. Projection Distances (CPD)

Analytical studies and empirical observations developed by Mansour and Juang [4] revealed that the major mismatch between clean and noisy LPC cepstral vectors, in the case of additive white noise, is the shrinkage of norms. Their first attempt was to compensate this effect by the incorporation of a scale factor λ into the standard cepstral Euclidean distance:

$$d_{P1} = (C_t - \lambda C_r)^T (C_t - \lambda C_r) \quad (3)$$

where C_t and C_r are the liftered cepstral column vector of the test and reference signals. From the orthogonality principle, they obtained that the optimal value of λ that minimizes d_{P1} is the projection of C_t onto C_r:

$$\lambda_{opt} = \frac{C_t^T C_r}{C_r^T C_r} \quad (4)$$

Other observations made in [4] were that cepstral vectors with higher norm are less affected than cepstral vectors with lower norm and that the angle between two cepstral vectors is less sensitive than the traditional Euclidean distance. These considerations leaded to propose a family of cepstral projection distances. The best results in [4] were obtained using:

$$d_{P2} = |C_t| (1 - \cos\beta) = |C_t| - \frac{C_t^T C_r}{|C_r|} \quad (5)$$

3. ONE-SIDED AUTOCORRELATION LINEAR PREDICTIVE CODING (OSALPC)

From the autocorrelation sequence R(n) we may define the one-sided (causal part of the) autocorrelation (OSA) sequence [6]

$$R^+(n) = \begin{cases} R(n) & n>0 \\ R(0)/2 & n=0 \\ 0 & n<0 \end{cases} \quad (6)$$

which verifies

$$R^+(n) + R^+(-n) = R(n), \quad -\infty \le n \le \infty \quad (7)$$

Its Fourier transform is the complex spectrum

$$S^+(\omega) = \frac{1}{2}\left[S(\omega) + jS_H(\omega)\right] \quad (8)$$

where $S(\omega)$ is the spectrum, 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 (8) and the analytic signal used in amplitude modulation, a spectral "envelope" $E(\omega)$ [7] can be defined as

$$E(\omega) = |S^+(\omega)| \quad (9)$$

This envelope characteristic, along with the high dynamic range of speech spectra, originate that $E(\omega)$ strongly enhances the highest power frequency bands. Thus, the noise components lying outside the enhanced frequency band are largely attenuated in $E(\omega)$ with respect to $S(\omega)$ (see Fig. 2).

Let us assume now that the speech signal x(n), whose autocorrelation is R(n), is given by the linear convolution

$$x(n) = h(n) * e(n) \quad (10)$$

where h(n) is the impulse response of a Mth-order all-pole filter driven by e(n), and e(n) is assumed to be a train of impulses for voiced sounds and white noise for unvoiced sounds. If $H(z) = 1/A(z)$ is the z-transform of h(n) and $S_e(\omega)$ is the power spectrum of e(n), it follows that

$$S(\omega) = \frac{S_e(\omega)}{|A(\omega)|^2} \quad (11)$$

The standard LPC approach performs a deconvolution of the speech signal since, assuming that $S_e(\omega)$ is a constant in (11), it obtains the characteristics of the vocal tract filter, H(z).

On the other hand, it is well known that $R^+(n)$ has the same poles than the signal [8]. Thus, if $B(\omega)$ is the Fourier transform of the driving function that obtains $R^+(n)$ at the output of the filter H(z), we can write

$$|S^+(\omega)| = \frac{|B(\omega)|}{|A(\omega)|} \quad (12)$$

In the same manner as LPC performs a linear prediction of the speech signal, we may consider a linear prediction of $R^+(n)$. This one-sided autocorrelation linear predictive coding (OSALPC) is equivalent to assume that $|B(\omega)|$ is constant in (12) and the square envelope, i.e. the spectrum of $R^+(n)$, is

$$E^2(\omega) = \frac{B}{|A(\omega)|^2} \quad (13)$$

Let us explore now the meaning of the above assumption. From (8) we can write $S(\omega)$ as a function of $A(\omega)$ and $B(\omega)$ as follows

$$S(\omega) = S^+(\omega) + (S^+(\omega))^* = \frac{B(\omega)}{A(\omega)} + \frac{B^*(\omega)}{A^*(\omega)} \quad (14)$$

and from identification of (14) and (11) it results that

$$S_e(\omega) = B(\omega) A^*(\omega) + B^*(\omega) A(\omega) \quad (15)$$

i.e. $B(\omega)$ depends on both $S_e(\omega)$ and $A(\omega)$ and can no longer be considered a constant. Thus, we can assert that the OSALPC technique does not actually perform a deconvolution between filter and excitation as does the LPC of the speech signal.

However, in spite of the OSALPC technique only

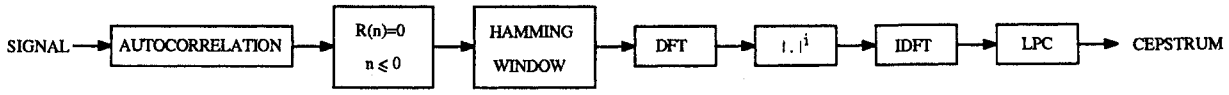
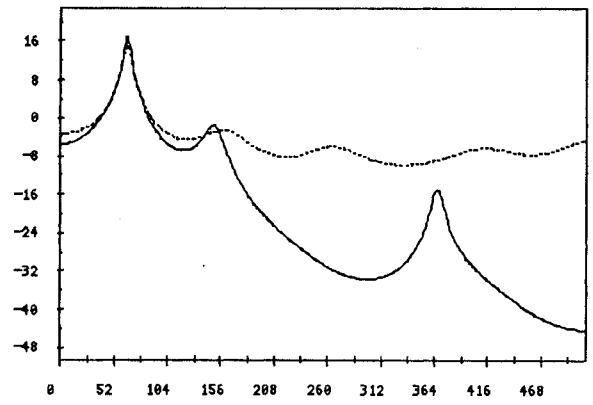


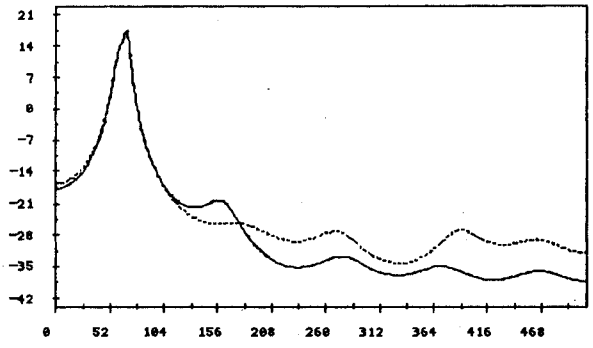
Fig.1 Block diagram for calculation of OSALPC (i =2) and SMC (i = 1) representations.

performs a partial deconvolution, as it will be seen its use in speech recognition achieves performances that are comparable to those obtained with the standard LPC approach for clean speech and outperforms it for noisy speech.

A block diagram to calculate the proposed technique is given in Fig. 1, with i = 2. In the next section, the proposed OSALPC technique will be compared with the SMC technique [5]. In fact, SMC follows the same scheme of Fig. 1 by setting i = 1.



(a) LPC spectrum



(b) OSALPC square envelope

Fig.2. Robustness of the OSALPC representation to additive white noise: (a) LPC spectrum and (b) OSALPC square envelope of a voiced speech frame in noisy free conditions (solid line) and SNR = 0 dB (dotted line).

4. SPEECH RECOGNITION EXPERIMENTS

This section reports the application of the techniques described above to recognize isolated words in a multispeaker task, with the HMM approach, in order to compare their performance and gain some perspective of the merit of the OSALPC representation in the presence of additive white noise.

4.1. Speech database and recognition system

The database used in our experiments consists of ten repetitions of the Catalan digits uttered by seven male and three female speakers (1000 words) and recorded in a quiet room.

The analog speech was first bandpass filtered to 100-3400 Hz. by an antialiasing filter and sampled at 8 KHz. The digitized clean speech was manually endpointed to determine the boundaries of each word. The endpoints obtained in this way were used in all our experiments. Clean speech was used for training in all the experiments. Noisy speech was simulated by adding zero mean white Gaussian noise to the clean signal so that the SNR of the resulting signal becomes ∞ (clean), 10 and 0 dB. No preemphasis was performed.

In the parameterization stage of the recognition system, the signal is divided into frames of 30 ms. at a rate of 15 ms. and each frame is characterized by its energy and L cepstral parameters obtained either by the standard LPC method or the others techniques based on the AR modeling in the autocorrelation domain. Before entering the recognition stage, the system evaluates the spectral difference with a time-average of 90 ms. [9]. In a similar way, the energy difference is calculated. The spectral vector and the spectral and energy differences are vector-quantized separately using three codebooks of 16 codewords (the size of the codebooks was optimized using Euclidean distance and the standard LPC technique). In that way, every frame of speech is represented by three independent symbols.

Each digit is characterized by a first order, left-to-right, discrete Markov model with 10 states. Training and testing are performed by the classical Baum-Welch and Viterbi algorithms, respectively. In all the recognition experiments the same whole data base was used for training the models and testing the recognition system.

4.2. Recognition results and conclusions

The first experiments carried out with the above described speech recognition system consisted of empirically optimizing the model order and the type of cepstral lifter in the standard LPC technique using the cepstral Euclidean distance. The preliminary recognition results showed that neither the model order nor the type of cepstral lifter are important for our task in noisy free conditions. However, in the presence of noise, a relatively higher order model is preferable. The best results were found for M = 12 and 14. These results appear in Table I for both cepstral lifters. It is clear from the table that the slope lifter outperforms the bandpass lifter.

Table I. Recognition rates for LPC representation and Euclidean distance

SNR (dB)		∞	10	0
M=12	Slope	100	97.3	68.8
	Bandpass	99.9	74.9	27.4
M=14	Slope	99.9	96.3	75.1
	Bandpass	99.9	91.7	40.5

In Table II, the results obtained by applying projection distance (CPD), concretely d_{P2} (5), upon the cepstral vector and Euclidean distance upon the difference cepstral vector are presented (with the same model orders and lifters). As can be seen, CPD performs as well as Euclidean distance for clean speech and obtains noticeably better results than it for noisy speech in the case of bandpass liftering. However, in the case of slope lifter CPD is better than Euclidean distance only for M = 12 (the recognition rates obtained in this case are better than those obtained with the model order and lifter used in [4]). It is worth noting that with CPD the difference between both lifters is not so drastic as with Euclidean distance.

Table II. Recognition rates for LPC representation and projection distance

SNR (dB)		∞	10	0
M=12	Slope	100	96.9	77.5
	Bandpass	100	94.9	70.7
M=14	Slope	99.8	94.8	75.6
	Bandpass	99.9	95.3	79.7

We also tested d_{P1} (3), but the results were slightly worse than those obtained with d_{P2}. Concretely, in the case of M = 12 and slope lifter, the recognition rates were 100, 96.2 and 73.0% for the three conditions tested. Finally, taking into account the fact that the difference cepstrum norm is also reduced in the case of additive white noise, we made a test applying d_{P2} in both cepstrum and difference cepstrum. The results were similar to those obtained in Table II: 100, 96.4 and 77.7 %, for M = 12 and slope lifter.

Finally, we tested the SMC and OSALPC representations in our speech recognition system. For the implementation of the OSALPC technique we used the classical biased autocorrelation estimator, while in the SMC implementation we used the unwindowed autocorrelation estimator as in [5]. The results are given in Table III using slope lifter and the same model orders that in Tables I and II. As occurs in the standard LPC approach, these model orders optimize the recognition rates.

Table III. Recognition rates for SMC and OSALPC representations and Euclidean distance.

SNR (dB)		∞	10	0
M=12	SMC	99.0	96.6	78.4
	OSALPC	98.8	95.9	79.6
M=14	SMC	99.0	96.0	78.5
	OSALPC	99.4	96.8	82.4

It is clear from the table that the performances of the OSALPC and SMC representations are similar for M = 12, but when the model order increases to 14 SMC scores become worse while OSALPC scores improve for the three conditions tested. These OSALPC results are the best of those obtained with all of the other techniques described in low SNR conditions.

ACKNOWLEDGEMENTS

The authors would like to thank Marcelino Zabalza and Jordi Rascado for their help in the software development.

REFERENCES

[1] F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Trans. on ASSP-23, nº 1, Feb. 1975, pp. 67-72.

[2] B.H. Juang, L.R.Rabiner and J.G. Wilpon, "On the use of band-pass liftering in speech recognition", IEEE Trans. on ASSP-35, nº 7, Jul. 1987, pp. 947-54.

[3] B.A. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise", IEEE Trans. on ASSP-35, nº 7, Jul. 1987, pp. 968-73.

[4] D. Mansour and B.H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition", IEEE Trans. on ASSP-37, nº 11, Nov. 1989, pp. 1959-71.

[5] D. Mansour and B.H. Juang, "The short-time modified coherence representation and its application for noisy speech recognition", IEEE Trans. on ASSP-37, nº 6, Jun. 1989, pp. 795-804.

[6] C. Nadeu, J. Pascual and J. Hernando, "Pitch determination using the cepstrum of the one-sided autocorrelation sequence", ICASSP'91, Toronto, May 1991, pp. 3677-80.

[7] M.A. Lagunas and M. Amengual, "Non-linear spectral estimation", ICASSP'87, Dallas, Apr. 1987, pp. 2035-38.

[8] D.P. McGinn and D.H. Johnson, "Reduction of all-pole parameter estimation bias by successive autocorrelation", ICASSP'83, Boston, Apr. 1983, pp.1088-91.

[9] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. on ASSP-34, Feb. 1986, pp. 52-59.