SPEECH RECOGNITION IN NOISY CAR ENVIRONMENT BASED ON OSALPC REPRESENTATION AND ROBUST SIMILARITY MEASURING TECHNIQUES*

Javier Hernando and Climent Nadeu

Signal Theory and Communications Department
Universitat Politècnica de Catalunya
08034 Barcelona, Spain
E-mail: javier@tsc.upc.es

ABSTRACT

The performance of the existing speech recognition systems degrades rapidly in the presence of background noise. The OSALPC (One-Sided Autocorrelation Linear Predictive Coding) representation of the speech signal has shown to be attractive for speech recognition because of its simplicity and its high recognition performance with respect to the standard LPC in severe conditions of additive white noise. The aim of this paper is twofold: 1) to show that OSALPC also achieves good performance in a case of real noisy speech (in a car environment), and 2) to explore its combination with several robust similarity measuring techniques, showing that its performance improves by using cepstral liftering, dynamic features and multilabeling.

1. INTRODUCTION

Speech recognition in noisy environments remains an unsolved problem even in the case of isolated word recognition with small vocabularies. In order to develop a system that operates robustly and reliably in the presence of noise, many techniques have been proposed in the literature [1] for reducing noise in each stage of the recognition process, particularly in feature extraction and similarity measuring.

A spectral estimation technique widely used in the parameterization stage of speech recognizers is Linear Predictive Coding (LPC) [2], which is equivalent to an AR modeling of the speech signal. Concretely, it has been shown that the use of the liftered LPC-cepstral coefficients in the standard Euclidean distance measure leads to the best results of those obtained with this model in both noise free [3] and noisy [4] conditions.

Recently, as an alternative representation of speech signals when noise is present, the authors proposed a parameterization technique called One-Sided Autocorrelation Linear Predictive Coding (OSALPC) [5] [6]. This technique, closely related with the Short-Time Modified Coherence (SMC) representation [7], is essentially an AR modeling of the causal part of the autocorrelation sequence and its use in noisy speech recognition is attractive because of its simplicity and its high recognition performance with respect to the standard LPC in severe conditions of additive white noise.

The aim of this paper is twofold: 1) to show that OSALPC also achieves good performance in a case of real noisy speech (in a car environment), and 2) to explore its combination with several robust similarity measuring techniques, showing that its performance improves by using cepstral liftering, dynamic features and multilabeling.

The paper is organized in the following way. In section 2 and 3 the OSALPC parameterization, and the robust similarity measuring techniques that are considered in this work are briefly revised (for more information see [8]). Section 4 is dedicated to show the experimental results obtained by applying these techniques, both separately and in combination, to recognize isolated words, in a multispeaker task, in real noisy car environment. Finally, in section 5 some conclusions are summarized from those results.

2. OSALPC REPRESENTATION

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

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

Its Fourier transform is the complex spectrum

$$S^+(\omega) = \frac{1}{2}[\, S(\omega) + jS_H(\omega)\,], \qquad (2)$$

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" $E(\omega)$ [9] can be defined as

$$E(\omega) = |S^+(\omega)|, \qquad (3)$$

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

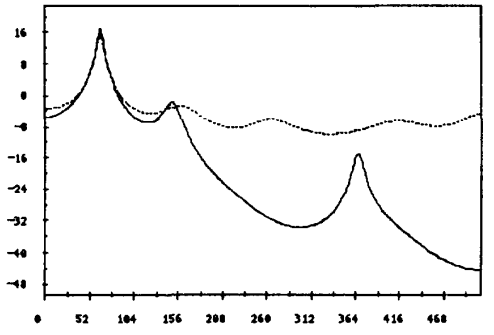
This envelope characteristic, along with the high dynamic range of speech spectra, originates that $E(\omega)$ strongly enhances the highest power frequency bands. Thus, the noise

* This work has been supported by the grant TIC 92-1026-C02/02

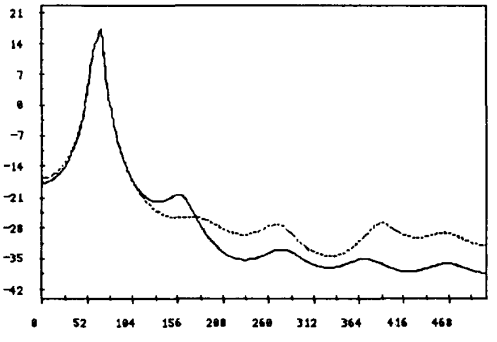
0-7803-1775-0/94 $3.00 © 1994 IEEE

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 the causal sequence $R^+(n)$ has the same poles than the signal itself [10]. 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 broad band noise. For this purpose, in the same manner as the standard LPC 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 $E^2(\omega)$. This is the basis of the OSALPC (One-Sided Autocorrelation Linear Predictive Coding) parameterization technique [5] [6]. The robustness of OSALPC to additive white noise is illustrated in Figure 1.



a) LPC spectrum



b) OSALPC square envelope

Fig. 1. Robustness of the OSALPC representation to additive white noise: a) LPC spectrum and b) OSALPC square envelope of a voiced speech frame in noise free conditions (solid line) and SNR equal to 0 dB (dotted line).

The Short-Time Modified Coherence (SMC) technique, proposed by D. Mansour and B.H. Juang [7], is also based on an AR modeling in the autocorrelation domain. However, whereas in the OSALPC technique the entries to the Levinson-Durbin algorithm (first p values of the autocorrelation of the one-sided autocorrelation sequence) are calculated from $R^+(n)$ using the classical biased autocorrelation estimator, in the SMC representation they are computed using a square root spectral shaper. In terms of the above OSALPC formulation, that difference actually consists of assuming in the SMC technique an all-pole spectral model for the envelope $E(\omega)$ instead of $E^2(\omega)$.

The OSALPC technique was compared in a previous work [5] [6] with both the standard LPC and the SMC technique, using speech signals that included additive white noise. In those tests, the OSALPC technique outperformed the other two for low SNR, using the conventional biased estimator to compute the one-sided autocorrelation. In the present investigation, OSALPC was implemented using the same one-sided autocorrelation estimator than SMC (i.e., the coherence estimator, which is defined in [7]), since we observed a slight improvement by using it instead of the biased estimator for the case of additive white noise. Actually, with the coherence estimator, the OSALPC representation achieved in our experiments better results than the SMC representation for every tested SNR, including clean speech [8].

3. ROBUST SIMILARITY MEASURING TECHNIQUES

3.1. Cepstral Lifter Optimization

In this paper three different cepstral lifters are considered:

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

Slope lifter: $w(n) = n$

Inverse of standard deviation lifter: $w(n) = \frac{1}{\sigma_{c(n)}}$ (4)

where $n = 1,...,L$ and $\sigma_{c(n)}$ is the standard deviation of the nth cepstral coefficient $c(n)$. If p denotes the prediction order, the value of L is typically $3p/2$ for the bandpass lifter [3] and p for the slope lifter [4] and the inverse of the standard deviation lifter [11].

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

3.2. Cepstral Projection Distance

Analytical derivatives and empirical observations developed by D. Mansour and B.H. Juang [12] revealed that the major mismatch between clean and noisy LPC-cepstral vectors, in the case of additive white noise, is the shrinkage of norms. They also observed 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. Those considerations led them to propose a family of cepstral projection distances for noisy speech recognition.

The best results [12] were obtained using:

$$dp = |\mathbf{C}_t| (1 - \cos\beta) = |\mathbf{C}_t| - \frac{\mathbf{C}_t^T \mathbf{C}_r}{|\mathbf{C}_r|}, \qquad (5)$$

where $\mathbf{C}_t$ and $\mathbf{C}_r$ are the liftered cepstral column vectors of the test and reference signals and β is the angle between them. This is the projection distance that will be used in the experimental results reported in section 4.

3.2. Dynamic Features

The so-called dynamic features are routinely used in current speech recognition systems in combination with short-term (static) spectral features. As their computation encompasses several adjacent frames, they are able to somewhat represent the time evolution of the spectrum of speech signals by providing smoothed estimates of the derivatives of the spectral parameter trajectories in the current frame.

In our work, we used the usual regression coefficient [13] that is applied to the cepstral sequence (delta-cepstrum) or the energy sequence (delta-energy). The window length, i.e. the number of frames used in the computation, was varied to get the best results.

3.4. Multilabeling

In the discrete HMM (DHMM) approach, for each incoming vector the quantizer makes a hard decision as to which of its codewords is the best match, and so the information about the degree to which the incoming vector matches other codewords is discarded. This information would be specially important in the case of noisy speech recognition, because that decision may be easily affected by the noise.

However, in the semicontinuous [14] HMM (SCHMM) and multilabeling [15] HMM (MLHMM) approaches, the vector quantizer makes a softer decision about which codeword is closest to the input vector, generating an output vector whose K components indicate the relative closeness to the K closest codewords. These components are estimated from the stochastic viewpoint in the SCHMM's and from the deterministic viewpoint in the MLHMM approach. In both cases, the recognition rates in noisy conditions are similar and outperform considerably those obtained using standard VQ [16]. Nevertheless, the MLHMM approach leads to algorithms that are more efficient than the SCHMM one. Because of this, the multilabeling method proposed in [16] will be used in the recognition experiments.

4. EXPERIMENTAL RESULTS

4.1. Database and Recognition System

The database used in noisy car environment experiments comes from the ESPRIT-ARS project and consists of 25 repetitions of the Italian digits uttered by 4 speakers, 2 males and 2 females, which were recorded in different noisy conditions: 5 repetitions with the engine and the fan off and 20 more with the engine on and different fan positions, 10 with the car stopped, 5 with the car running at 70 km/h and 5 with the car running at 130 km/h. The system was trained with the signals uttered when the engine and the fan were off, i.e., in noise free conditions, and in the test phase the noisy signals were used.

In the parameterization stage, the speech signal, sampled at 8 kHz, quantized using 12 bits per sample, manually endpointed and preemphasized, was divided into frames of 30 ms at a rate of 15 ms and each frame was characterized by its liftered cepstral parameters, obtained either by the standard LPC method or the new OSALPC technique. In some tests the dynamic parameters of the frame were also obtained. Each information was separately vector-quantized using a codebook of 64 codewords by means of standard VQ or the multilabeling method with either the standard Euclidean distance or the new cepstral projection distortion measure. Each digit was characterized by a first order, left-to-right, Markov model of 10 states without skips. Training and testing were performed using Baum-Welch and Viterbi algorithms, respectively.

4.2. Recognition Results

The first experiments carried out with the above described speech recognition system consisted of empirically optimizing the prediction order and the type of cepstral lifter using the standard cepstral Euclidean distance upon the static cepstrum and standard VQ. Preliminary results showed that neither the prediction order nor the type of cepstral lifter are important for our task in noise free conditions. However, in the presence of noise the recognition results are very sensitive to both factors. The best results were obtained using prediction order equals to 16 and inverse of the standard deviation lifter for the standard LPC parameterization and slope lifter for the new OSALPC technique, i.e. a relatively high prediction order and one of the two non-symmetrical cepstral lifters.

Actually, a relatively high value of the prediction order can provide more robust estimates of the autocorrelation in the presence of broad-band noise due to fact that the sensitivity to this type of noise tends to decrease along the autocorrelation index.

In Figure 2, the recognition rates obtained using these optimum orders and lifters are compared, in terms of the car speed, with those obtained using an order equal to 8 and bandpass lifter in noise free conditions. Notice that the results are very sensitive to those factors, and that relatively high prediction orders and non-symmetrical cepstral lifters are preferable in noisy conditions. It can also be seen that, using the optimum orders and lifters, OSALPC noticeably outperforms LPC in severe noisy conditions.

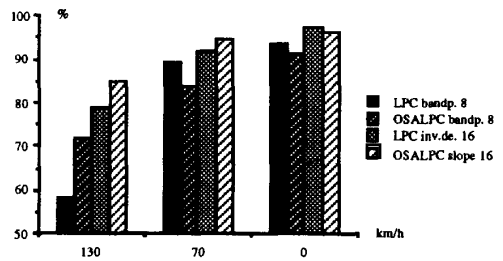


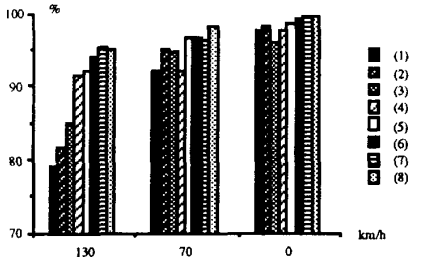
Fig. 2. Optimization of prediction order and cepstral liftering in LPC and OSALPC techniques

The results obtained using cepstral projection distance were not better than those obtained applying the standard Euclidean distance. The type of noise considered in this work may justify these results, since the cepstral projection distance measure was proposed for the case of white noise.

Regarding to dynamic features, the use of delta-cepstrum and delta-energy, in the case of the standard LPC parameterization, and the use of delta-cepstrum, in the case of the OSALPC technique, provided the best results. The best results were obtained using a window length of 240 ms for the estimation of delta-parameters.

Excellent results were also obtained applying the multilabeling method instead of the standard VQ approach. The tradeoff between computational load and recognition accuracy led us to consider only the information corresponding to the five codewords closest to the incoming vector.

The combination of these techniques, excepting cepstral projection distance measure, provided even better results than those obtained applying each technique separately. In Figure 3, recognition rates obtained using the optimum orders and lifters are compared in terms of the parameterization -LPC or OSALPC- and vector quantization -standard VQ or multilabeling (ML)- employed and either using or not dynamic features. The various combinations of techniques have been ordered taking into account the recognition rates obtained in severe noisy conditions.



(1) LPC. Standard VQ. Cepstrum
(2) LPC. ML. Cepstrum
(3) OSALPC. Standard VQ. Cepstrum
(4) OSALPC. ML. Cepstrum
(5) LPC. Standard VQ. Cepstrum, delta-cepstrum and delta-energy
(6) LPC. ML. Cepstrum, delta-cepstrum and delta-energy
(7) OSALPC. Standard VQ. Cepstrum and delta-cepstrum
(8) OSALPC. ML. Cepstrum and delta-cepstrum

Fig. 3. Comparison and combination of techniques

As it can be observed in Figure 3, the OSALPC technique without using delta-cepstrum obtains excellent results in severe noisy conditions, but the standard LPC technique results are better than OSALPC results in almost noise free conditions. However, using delta-cepstrum, OSALPC outperforms LPC in all the considered conditions. On the other hand, it can be seen that the multilabeling method yields excellent results combined with the use of energy and dynamic information. The best results are obtained using OSALPC parameterization, delta-cepstrum and multilabeling.

5. CONCLUSIONS

From the application of the OSALPC parameterization on a system based on the Hidden Markov Modeling and Vector Quantization approaches for speech recognition in real noisy car environment in combination with several robust similarity measuring techniques, some conclusions can be summarized:

a) When linear prediction techniques are used in the parameterization stage, a relatively high prediction order and the use of a non-symmetrical lifter are preferable.

b) Cepstral representation based on linear prediction of the one-sided autocorrelation sequence (OSALPC) provides excellent results in severe noisy conditions.

c) The cepstral projection distance measure does not yield good results in this noisy car environment.

d) The addition of dynamic features is very useful in all the considered conditions.

e) The multilabeling technique noticeably outperforms the standard VQ method.

f) The combination of those techniques, excepting the cepstral projection distortion measure, provides better results than those obtained applying each technique separately.

ACKNOWLEDGMENTS

The authors would like to thank Joan Dachs for his help in software development.

REFERENCES

[1] B.H. Juang, Computer Speech and Language, vol. 5, pp. 275-294, 1991.
[2] F. Itakura, IEEE Trans. ASSP, vol. 23, pp. 67-72, 1975.
[3] B. H. Juang, L.R. Rabiner and J.G. Wilpon, IEEE Trans. ASSP, vol. 35, pp. 947-954, 1987.
[4] B.A. Hanson and H. Wakita, IEEE Trans. ASSP, vol. 35, pp. 968-973, 1987.
[5] J. Hernando and C. Nadeu, Proc. EUROSPEECH'91, Genova, pp. 91-94, 1991.
[6] J. Hernando, C. Nadeu and E. Lleida, Proc. ICSLP'92, Banff, pp. 1593-1996, 1992.
[7] D. Mansour and B.H. Juang, IEEE Trans. ASSP, vol. 37, pp. 795-804, 1989.
[8] J. Hernando, Ph.D. Dissertation, Dpt. Signal Theory and Communications, Polytechnical University of Catalonia, Barcelona 1993.
[9] M.A. Lagunas and M. Amengual, Proc. ICASSP'87, Dallas, pp. 2035-2038, 1987.
[10] D.P. McGuinn, D.H. Johnson, Proc. ICASSP'83, Boston, pp. 1088-1091, 1983.
[11] Y. Tohkura, IEEE Trans. ASSP, vol. 35, pp. 1414-1422, 1987.
[12] D. Mansour and B.H. Juang, IEEE Trans. ASSP, vol. 37, pp. 1959-1971, 1989.
[13] S. Furui, IEEE Trans. ASSP, vol. 34, pp. 52-59, 1986.
[14] X.D. Huang, IEEE Trans. ASSP, vol. 40, pp. 1062-1067, 1992.
[15] M. Nishimura and K. Toshioka, Proc. ICASSP'87, Dallas, pp. 1163-1166, 1987.
[16] J. Hernando, J.B. Mariño, and C. Nadeu, Proc. EUROSPEECH'93, Berlin, pp. 1643-1646, 1993.