# A COMPARATIVE STUDY OF TECHNIQUES FOR HMM-BASED SPEECH RECOGNITION IN NOISY CAR ENVIRONMENT

J. Hernando, C. Nadeu, J.B. Mariño
Dep. of Signal Theory and Communications (U.P.C.)
Ap. 30002, 08071 Barcelona, Spain
javier@tsc.upc.es

## ABSTRACT

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. The aim of this paper is to report the application of several robust techniques on a system based on the HMM (Hidden Markov Models) and VQ (Vector Quantization) approaches for speech recognition in noisy car environment: parameterization based on the linear prediction of the causal part of the autocorrelation sequence (OSALPC) -proposed by the authors in [1] [2]-, optimization of spectral model order and cepstral lifter, cepstral projection distance measure, dynamic information 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 [3] 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) [4], based on an AR modeling of the speech signal. Concretely, it has been shown that the use of the LPC-cepstral coefficients liftered in the standard Euclidean distance measure lead to the best results of those obtained with this model in both noise free [5] and noisy [6] conditions. Furthermore, it is well known that the use of a relatively high prediction order can provide more robust estimations.

Recently, as an alternative representation of speech signals when noise is present, the authors proposed a parameterization technique called OSALPC (One-Sided Autocorrelation Linear Predictive Coding) [1], based on an AR modeling of the causal part of the speech autocorrelation sequence. This technique, closely related with the SMC representation [7], is interesting in noisy speech recognition because of its simplicity and computational efficiency and outperforms the standard LPC approach in speech recognition in severe noisy conditions.

Furthermore, there is no obvious reason to maintain the symmetry characteristics of the Euclidean distance if one knows the reference and test signals have different degree of noisy corruption. This is the basis of the cepstral projection distance measures proposed in [8], which take into account the effects of additive white noise upon the LPC-cepstral representation.

On the other hand, it is well known that dynamic features play an important role in the human speech perception and are more robust to noise and to interspeaker variability than static features [9]. So static and dynamic information are used in most of the existing speech recognition systems.

Applying multilabeling techniques [10] to quantize both kinds of information in a speech recognition system based on the HMM (Hidden Markov Models) and VQ (Vector Quantization) approaches, the authors [11] have obtained excellent results in noisy speech recognition compared with those obtained using standard VQ techniques.

The aim of this paper is to make a comparative study of these techniques in noisy speech recognition. In section 2 all the robust techniques mentioned above are briefly revised (for more information see [12]). Section 3 reports the application of these techniques to recognize isolated words in a multispeaker task using a system based on the HMM and VQ approaches in noisy car environment. Finally, in section 4 some conclusions are summarized.

## 2. ROBUST RECOGNITION TECHNIQUES

### 2.1. Prediction order and cepstral lifter optimization

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 smoothing in noisy conditions.

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

In the case of broad-band noise, lower order autocorrelation and cepstral coefficients are more affected by the noise than higher order ones. Then it is suggested that it would be preferable to use a relatively high prediction order and an increasing lifter with the quefrency.

### 2.2. One-Sided Autocorrelation Linear Predictive Coding (OSALPC)

From the autocorrelation sequence R(n) we may define the one-sided autocorrelation sequence $R^+(n)$ as its causal part (i.e., R(n) is twice the even part or $R^+(n)$)**.** The real part of its Fourier transform $S^+(\omega)$ is the spectrum $S(\omega)$, i.e. the Fourier transform of R(n), an the imaginary part is the Hilbert transform of the $S(\omega)$. Due to the analogy between $S^+(\omega)$ and the analytic signal used in amplitude modulation, a spectral "envelope" $E(\omega)$ [12] can be defined as its magnitude. 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 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.

It is then suggested that the AR parameters of the signal can be more reliably estimated using linear prediction techniques from $R^+(n)$ than directly from the signal itself when it is corrupted by noise. This is the basis of the OSALPC (One-Sided Autocorrelation Linear Predictive Coding) parameterization technique proposed by the authors in [1] as a robust representation of speech signal when noise is present.

### 2.3. Cepstral projection distance

Analytical studies and empirical observations developed in [8] revealed that additive white noise reduces the norm of LPC-cepstral vectors and that the angle between two cepstral vectors is minimally sensitive to this corruption. That study led to the proposition of a family of distances and the best results were obtained using the projection between the test cepstral vector and the reference normalized cepstral vector.

### 2.4. Dynamic information

The first-order finite difference is intrinsically noisy. To alleviate this problem, Furui [9] suggested the application of regression analysis to each time function of the cepstrum coefficients (delta-cepstrum) and the log-energy (delta-energy) over a finite length window. That is the estimation of the dynamic information used in this paper.

There are two main strategies for the addition of these informations in a speech recognition system based on the VQ and HMM approaches: composite distance (a supervector is built weighting each component and a unique symbol is obtained from the VQ stage) and multiple codebooks (each information is quantized separately and is considered statistically independent in the HMM framework). Preliminary experiments led us to use this latter approach in the recognition results reported in this paper.

### 2.4. Multilabeling

In the discrete HMM (DHMM) approach, for each incoming vector the VQ 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 hard decision can be easily modified by the noise added to the speech.

However, in the semicontinuous HMM (SCHMM) and multilabeling HMM (MLHMM) [11] approaches, the VQ makes a soft decision about which codeword is closest to the input vector, generating an output vector whose K components indicate the relative closeness of the K closest codewords. These components are estimated from the stochastic viewpoint in the SCHHM'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. Nevertheless, the MLHMM approach is more computationally efficient that the SCHMM approach. Because of this, that latter approach will be used in the recognition experiments.

## 3. EXPERIMENTAL RESULTS

### 3.1. Database and recognizer setup

The database used in our experiments is from the ESPRIT-ARS project and consists of 25 repetitions of the Italian digits uttered by 4 speakers, 2 males and 2 females, inside a car 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 signal uttered with the engine and the fan off, i.e., in noise free conditions, and in the test phase the noisy signals were used. In both phases, the signals were manually endpointed.

In the parameterization stage, the speech signal, sampled at 8 kHz, quantized using 12 bits per sample and preenfasized, 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 or the new OSALPC techniques. In some tests the log-energy and the dynamic parameters of the frame were also obtained. Each information was vector-quantized separately by means a codebook of 64 codewords, using the standard VQ or the multilabeling approaches and the Euclidean or projection distances.

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.

### 3.2. Recognition results

The first experiments carried out with the above described speech recognitions system consisted of empirically optimizing the prediction order and the type of cepstral lifter (rectangular, bandpass, slope and inverse of the standard deviation), using cepstral Euclidean distance, only static cepstrum and standard VQ. 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. Furthermore, it has been obseved that the OSALPC technique is less sensitive than the standard LPC approach to changes in the prediction order and cepstral lifter.

In figure 1, the recognition rates obtained, in function of the car speed, using these optimum orders and lifters are compared with those obtained using the most used order (8, for sampling frequency of 8 kHz) and lifter (bandpass) in noise free conditions. It can be seen that the results are very sensitive to both parameters of the system and that relatively high prediction orders and increasing cepstral lifters are preferable in noisy conditions. Also it can bee seen that, using the optimum orders and lifters, OSALPC outperforms noticeably LPC in severe noisy conditions.

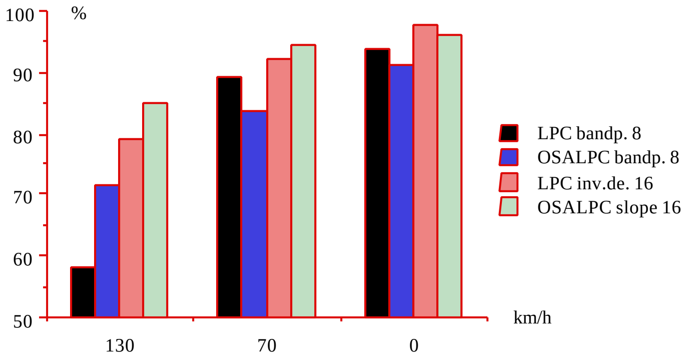


Figure 1

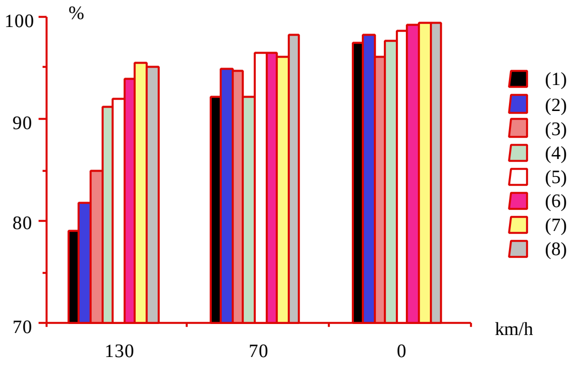
The results obtained using cepstral projection distance were worse than those obtained applying the standard Euclidean distance. The type of the noise considered in this work can justify these results, so the cepstral projection distance measure was proposed in the case of white noise.

With respect to the addition of energy and dynamic information, the use of delta-cepstrum and delta-energy, in the case of LPC parameterization, and the use of delta-cepstrum, in the case of OSALPC technique, provides excellent results. The best results were obtained using a window length of 240 ms for the estimation of the regressive parameters.

Also excellent results are 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.

Combination of these techniques, except the cepstral projection distance measure, provides even better results than those obtained applying each technique separately. In figure 2 recognition rates obtained, using the optimum orders and lifters, are compared in function of the parameterization -LPC or OSALPC- and vectorial quantization - standard or multilabeling- employed and the use or not of the energy and dynamic information. Different combinations of techniques have been ordered taking into account the recognition rates obtained in severe noisy conditions.

It can be seen in this figure that 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.



(1) LPC, standard VQ, only ceps.
(2) LPC, ML, only ceps.
(3) OSALPC, standard VQ, only ceps.
(4) OSALPC, ML, only ceps.
(5) LPC, standard VQ, several inf.
(6) LPC, ML, several inf .
(7) OSALPC, standard VQ, several inf.
(8) OSALPC, ML, several inf

Figure 2

## 4. CONCLUSIONS

Several robust speech recognition techniques in noisy environments, particularly, in feature extraction and similarity measuring, have been revised in this paper. From a comparative study of these techniques using a system based on the HMM (Hidden Markov Models) and VQ (Vector Quantization) approaches to recognize isolated words in a multispeaker task in noisy car environment, some conclusions can be summarized:

a) When linear prediction techniques are used in the parameterization stage, it is preferable a relatively high order prediction and the use of an increasing lifter with the quefrency.

b) The cepstral projection distance measure does not yield good results in the case of the type of noise considered.

c) Cepstral representation based on the linear prediction of the one-sided autocorrelation sequence (OSALPC), proposed by the authors in [1], provides excellent results in severe noisy conditions. Furthermore, it has been seen that the OSALPC technique is less sensitive than the standard LPC approach to changes in the prediction order and cepstral lifter.

d) The addition of energy and dynamic information is very useful in all the conditions considered.

e) Multilabeling method otuperforms noticeably standard vector quantization.

f) Combination of these techniques, except the cepstral projection distortion measure, provides even better results than those obtained applying each technique separately.

## REFERENCES

[1] J. Hernando, C. Nadeu, Proc. EUROSPEECH'91, Genova, pp. 91-94.
[2] J. Hernando, C. Nadeu, E. Lleida, Proc. ICSLP'92, Banff (Alberta, Canada), pp. 1593-96.
[3] B.H. Juang, Computer Speech and Language, vol. 5, 1991, pp. 275-94.
[4] F. Itakura, IEEE Trans. ASSP, vol. 23, 1975, pp. 67-72.
[5] B. H. Juang, L.R. Rabiner, J.G. Wilpon, IEEE Trans. ASSP, vol. 35, 1987, pp. 947-54.
[6] B.A. Hanson, H. Wakita, IEEE Trans. ASSP, vol. 35, 1987, pp. 968-73.
[7] D. Mansour, B.H. Juang, IEEE Trans. ASSP, vol. 37, 1989, pp. 795-804.
[8] D. Mansour, B.H. Juang, IEEE Trans. ASSP, vol. 37, 1989, pp. 1959-71.
[9] S. Furui, IEEE Trans. ASSP, vol. 34, 1986, pp. 52-59.

[10] M. Nishimura, K. Toshioka, Proc. ICASSP'87, Dallas, pp. 1163-66.
[11] J. Hernando, J.B. Mariño y C.Nadeu, EUROSPEECH´93, Berlin.
[12] J. Hernando, Ph.D. Disertation, Dpt. Signal Theory and Communications,UPC, Barcelona, Mayo 1993.
[13] M.A. Lagunas, M. Amengual, Proc. ICASSP-87, Dallas, pp. 2035-38.