ELSEVIER

Speech Communication 21 (1997) 17–31

SPEECH COMMUNICATION

Speech recognition in a noisy car environment based on LP of the one-sided autocorrelation sequence and robust similarity measuring techniques [1]

Javier Hernando *, Climent Nadeu, José B. Mariño

Department of Signal Theory and Communications, Polytechnical University of Catalonia, Barcelona, Spain

Received 16 January 1995; revised 19 August 1996

Abstract

The performance of the existing speech recognition systems degrades rapidly in the presence of background noise. A novel representation of the speech signal, which is based on Linear Prediction of the One-Sided Autocorrelation sequence (OSALPC), has shown to be attractive for noisy speech recognition because of both its high recognition performance with respect to the conventional LPC in severe conditions of additive white noise and its computational simplicity. The aim of this work is twofold: (1) to show that OSALPC also achieves a 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.

Résumé

Les performances des systèmes actuels de reconnaissance de parole se dégradent rapidement en présence de bruit. Une nouvelle représentation du signal de parole, basée sur la prédiction linéaire de séquence d'autocorrélation unilatérale (One-Sided Autocorrelation Linear Prediction: OSALPC), s'est avérée être intéressante pour la reconnaissance de la parole bruitée, à la fois pour ses bonnes performances (par rapport au codage LPC conventionnel) dans des conditions difficiles de bruit blanc additif et pour sa simplicité de calcul. Le but du travail présenté dans cet article est double: (1) il s'agit de montrer que OSALPC fournit également de bonnes performances pour de la parole bruitée en contexte réel d'usage (en voiture), et (2) d'explorer sa combinaison avec diverses techniques robustes de mesure de similarité, en montrant que ses performances s'améliorent en utilisant une pondération cepstrale, des indices dynamiques et l'étiquetage multiple.

Keywords: Speech recognition; Noise robustness; Feature extraction; Spectral analysis of speech; Distortion measures; Vector quantization

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

* Corresponding author. E-mail: javier@gps.tsc.upc.es.

[1] This work has been supported by the grants TIC-92-0800-C05/04 and TIC-92-1026-C02/02.

PII S0167-6393(96)00074-X

presence of noise, many techniques have been proposed in the literature (Juang, 1991) for reducing noise in each stage of the recognition process, particularly in feature extraction and similarity measuring.

Regarding the parameterization stage, a spectral estimation technique widely used in speech recognition is Linear Predictive Coding (LPC) (Itakura, 1975), which is equivalent to AR modeling of the speech signal. Concretely, it has been shown that the use of the liftered LPC-cepstral coefficients in the conventional Euclidean distance measure leads to the best results of those obtained with this model (Juang et al., 1987). However, the conventional LPC technique is known to be very sensitive to the presence of additive noise. This fact leads to poor recognition rates when this technique is used in speech recognition under noisy conditions, even if only a moderate level of contamination is present in the speech signal.

Linear prediction of the autocorrelation sequence has been the common approach to several robust spectral estimation methods for noisy signals presented in the past. For speech recognition, Mansour and Juang (1989a) proposed the Short-time Modified Coherence (SMC) as a robust representation of speech based on that approach. On the other hand, Cadzow (1982) introduced the use of an overdetermined set of Yule–Walker equations for robust modeling of time series. Although Cadzow applies linear prediction to the signal, his method can also be interpreted as performing linear prediction in the autocorrelation domain. Both methods rely, either explicitly or implicitly, on the fact that the autocorrelation sequence is less affected by broad-band noise than the signal itself, especially at high lag indices.

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) (Hernando et al., 1992). In this work, the causal part of the autocorrelation sequence and its mathematical properties are considered. As this sequence shares its poles with the signal $x(n)$, it provides a good starting point for LPC modeling. Also, it is closely related to the SMC representation and Cadzow's method. All of them can be interpreted as AR modeling of either a spectral function named "envelope" or its square. This interpretation, which is based on the properties of the one-sided autocorrelation, provides more insight into the various methods. As shown in (Hernando et al., 1992), the use of OSALPC in noisy speech recognition is attractive because of both its high recognition performance with respect to conventional LPC in severe conditions of additive white noise and its computational simplicity.

The aim of this work 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 filtering of spectral parameters – cepstral liftering (Juang et al., 1987; Hanson and Wakita, 1987; Tohkura, 1987) and differential parameters (Furui, 1986) – and multilabeling (Hernando et al., 1993).

The paper is organized in the following way. In Sections 2 and 3 the OSALPC parameterization and the robust similarity measuring techniques that are considered in this work are briefly reviewed (for more information see (Hernando, 1993)). 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 real noisy car environment, in a multispeaker task, using a speech recognition system based on the Hidden Markov Model (HMM) and Vector Quantization (VQ) approaches and trained in clean conditions. Finally, in Section 5 some conclusions are summarized from those results.

2. OSALPC representation

From the autocorrelation sequence $R(m)$, we define the one-sided autocorrelation (OSA) sequence $R^+(m)$ as its causal part in the following way:

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

which verifies

$$R(m) = R^{+}(m) + R^{+}(-m), \quad -\infty \leq m \leq \infty. \tag{2}$$

Its Fourier transform is the complex "spectrum"

$$S^{+}(\omega) = \tfrac{1}{2}[S(\omega) + \mathrm{j}S_{\mathrm{H}}(\omega)], \tag{3}$$

where $S(\omega)$ is the real spectrum of the signal, i.e. the Fourier transform of $R(m)$, and $S_{\mathrm{H}}(\omega)$ is the Hilbert transform of $S(\omega)$.

Due to the analogy between $S^{+}(\omega)$ in Eq. (3) and the analytic signal used in amplitude modulation, a spectral "envelope" $E(\omega)$ (Lagunas and Amengual, 1987) can be defined as

$$E(\omega) = |S^{+}(\omega)|, \tag{4}$$

whose square, the squared envelope, is the spectrum of the OSA sequence.

Fig. 1 shows that the speech spectral envelope $E(\omega)$ strongly enhances the highest power frequency bands with respect to $S(\omega)$. This may be due to the large dynamic range of speech spectra. Consequently, in noisy speech signals the noise components lying outside the enhanced frequency bands are largely attenuated in $E(\omega)$ with respect to $S(\omega)$, and thus $E(\omega)$ is more robust to broad-band noise than $S(\omega)$. This robustness to broad-band noise can be viewed in the time domain: the OSA sequence, whose spectrum is the squared spectral envelope $E^{2}(\omega)$, is less affected by broad-band noise than the signal itself, especially at high lag indices. This idea of enhancing spectral peaks relative to spectral valleys has been already used in order to improve robustness to noisy conditions: weighted distortion measures (Matsumoto and Imai, 1986), root cepstral analysis (Alexandre and Lockwood, 1993), etc.

On the other hand, as is well known, the OSA sequence $R^{+}(m)$ and the signal $x(n)$ have the same poles (McGinn and Johnson, 1983). Those two properties, i.e. robustness to broad-band noise and pole preservation, suggest that AR parameters of the speech signal can be more reliably estimated from the OSA sequence than directly from the signal $x(n)$ when $x(n)$ is corrupted by broad-band noise. Thus, as the conventional LPC technique assumes an all-pole model of the speech spectrum $S(\omega)$, we may apply linear prediction to the OSA

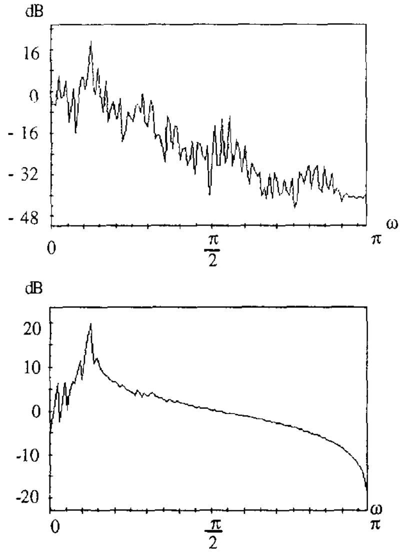


Fig. 1. Spectrum (top graph) adn its squared envelope (bottom graph) of a voiced speech frame in noise free conditions.

sequence, assuming 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 proposed in (Hernando et al., 1992) as a robust representation of speech signal when noise is present.

A straightforward algorithm is proposed to calculate the OSALPC cepstrum coefficients, that consists in applying the (windowed) autocorrelation method of linear prediction to an estimation of the OSA sequence (see block diagram in Fig. 2):

(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 consider the well known tradeoff between variance and frequency resolution of the spectral estimate).

(b) Secondly, the Hamming window from $m = 0$ to M is applied to such estimated OSA sequence.

(c) Thirdly, if p is the prediction order, the first $p + 1$ autocorrelation values of that OSA sequence are computed from $m = 0$ to p using the conventional biased estimator, i.e. the one that is commonly employed in speech processing.

(d) Then these values are used as entries to the Levinson–Durbin algorithm to estimate AR parameters a_k, $k = 1, \ldots, p$.

(e) Finally, the cepstral coefficients corresponding to the model are recurrently computed from those AR parameters.

The robustness of OSALPC to additive white noise is illustrated in Fig. 3. As can be seen in this figure, the OSALPC squared envelope shows a prominent first formant and its whole curve is more robust to additive white noise than that of the LPC spectrum. In this case, the conventional biased autocorrelation estimator was used to compute the OSA sequence from the signal.

Fig. 3 also shows that spurious peaks may appear in the OSALPC squared envelope. Probably, they are due to the fact that the OSALPC technique performs only a partial deconvolution of the speech signal. However, in spite of that, it shows a better speech recognition performance than conventional LPC in severe conditions of additive white noise (Hernando et al., 1992).

The Short-Time Modified Coherence (SMC) technique (Mansour and Juang, 1989a) is also based on AR modeling in the autocorrelation domain. However, whereas in the OSALPC technique the entries to the

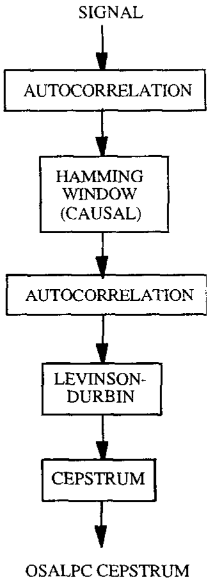


Fig. 2. Block diagram for the calculation of the OSALPC cepstrum.

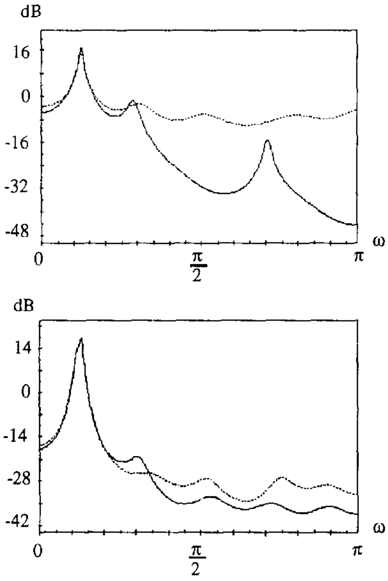


Fig. 3. Robustness of the OSALPC represetnation to additive white noise: LPC spectrum (top graph) and OSALPC squared envelope (bottom graph) of a voiced speech frame in noise free conditions (solid line) and SNR equal to 0 dB (dotted line).

Levinson–Durbin algorithm (the first p values of the autocorrelation of the OSA sequence) are calculated from the OSA sequence using the classical biased autocorrelation estimator, in the SMC representation they are computed using a square root spectral shaper. In fact, in terms of the above formulation, that difference lies in assuming in the SMC technique an all-pole spectral model for the envelope $E(\omega)$ instead of $E^2(\omega)$.

On the other hand, the name of the Short-Time Modified Coherence representation derives from the usage of a particular estimator, which is referred to as coherence in (Mansour and Juang, 1989a), to compute the OSA sequence from the signal. This estimator is a more homogeneous measure than the conventional biased autocorrelation estimator in the sense that every estimated value is computed using the same number of signal samples, whereas in the conventional estimator the number of signal samples employed to estimate $R(m)$ decreases along the index m. That property does not have much relevance in the estimation of the autocorrelation entries to the Levinson–Durbin, since only the first $p+1$ values are considered and usually $p \ll N$. However, it may be important in the estimation of the OSA sequence from the speech signal since the OSA length considered in both OSALPC and SMC techniques is $M = N/2$, not negligible with respect to N.

The OSALPC technique was compared in a previous work (Hernando et al., 1992) with both the conventional LPC and the SMC techniques, 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 OSA sequence from the signal. In the present work, OSALPC was implemented using the coherence estimator, since we observed a slight improvement by using it instead of the biased estimator for clean speech and moderate levels 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 (Hernando, 1993).

On the other hand, OSALPC is also closely related to the overdetermined set of Yule–Walker equations proposed by Cadzow (1982) to seek ARMA models of time series. Since an AR(p) process contaminated by additive white noise becomes an ARMA(p,p) process with the same poles as the AR(p) process, Cadzow's method can be used to estimate the parameters of this noisy AR process, just by setting the same AR and MA orders in the so called Least Squares Modified Yule–Walker Equations (LSMYWE) (Marple, 1987).

The relationship between OSALPC and LSMYWE techniques is illustrated by the matrix equation

$$
\text{OSALPC}\left\{\begin{bmatrix}
R(1) & 0 & 0 & \cdots & 0 \\
R(2) & R(1) & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
R(p) & R(p-1) & R(p-2) & \cdots & 0 \\
R(p+1) & R(p) & R(p-1) & \cdots & R(1) \\
R(p+2) & R(p+1) & R(p) & \cdots & R(2) \\
\vdots & \vdots & \vdots & & \vdots \\
R(M) & R(M-1) & R(M-2) & \cdots & R(M-p) \\
0 & R(M) & R(M-1) & \cdots & R(M-p+1) \\
0 & 0 & R(M) & \cdots & R(M-p+2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & R(M)
\end{bmatrix}\right.
\begin{bmatrix} 1 \\ a_1 \\ a_2 \\ \cdots \\ a_p \end{bmatrix}
$$

$$
= \begin{bmatrix}
e(1) \\ e(2) \\ \vdots \\ e(p) \\ e(p+1) \\ e(p+2) \\ \vdots \\ e(M) \\ e(M+1) \\ e(M+2) \\ \vdots \\ e(M+p)
\end{bmatrix}
\begin{matrix}
\vphantom{e(1)} \\ \vphantom{e(2)} \\ \vphantom{\vdots} \\ \vphantom{e(p)} \\
\left.\vphantom{\begin{matrix} e(p+1) \\ e(p+2) \\ \vdots \\ e(M) \end{matrix}}\right\}\text{LSMYWE,} \\
\vphantom{e(M+1)} \\ \vphantom{e(M+2)} \\ \vphantom{\vdots} \\ \vphantom{e(M+p)}
\end{matrix}
\tag{5}
$$

where M denotes the highest autocorrelation lag used and $e(m)$ is the error to be minimized. The minimization of the norm of the full error vector $\{e(m)\}_{m=1,\dots,M+p}$ with respect to the AR parameters a_k is equivalent to the application of the (windowed) autocorrelation method of linear prediction to the sequence $R(m)$, $m=1,\dots,M$, that is the OSALPC technique. On the other hand, the LSMYWE technique minimizes the norm of the subvector $\{e(m)\}_{m=p+1,\dots,M}$ and so it amounts to applying the (unwindowed) covariance method of linear prediction upon the same range of autocorrelation lags. When $M=2p$, LSMYWE are the Modified Yule–Walke Equations (Marple, 1987) for an ARMA(p,p) process. In both OSALPC and LSMYWE, only autocorrelation lags corresponding to the OSA sequence are employed. The only difference between both techniques is the range of error values considered in the minimization.

In spite of the similarity between all these techniques, the OSALPC representation outperforms the LSMYWE and SMC techniques in speech recognition for low SNRs of additive white noise (Hernando et al., 1992). On the other hand, as far as the computational complexity of the algorithms is concerned, OSALPC and SMC techniques are much more efficient than the LSMYWE technique because they use the Levinson–Durbin algorithm.

Finally, it is worth noting that the OSALPC technique may be included in the field of higher-order spectral estimation, due to the fact that the squared envelope $E^2(\omega)$ is the Fourier transform of the autocorrelation of the OSA sequence, that is a particular fourth-order moment of the signal.

3. Robust similarity measuring techniques

This section is devoted to briefly review several robust similarity measuring techniques that are used in the recognition experiments reported in Section 4. Herewith, the term similarity measuring techniques is used in a wide sense; it not only includes distance measure techniques between vectors of cepstral parameters, as the Euclidean or the projection distance (Section 3.3), but also other related techniques which are also used in the comparison stage of the recognition system, as filtering of spectral parameters (Section 3.1) or multilabeling (Section 3.2).

3.1. Filtering spectral parameters

The parametric representation of a speech utterance consists of a sequence of vectors, one per frame, whose components usually are some kind of cepstral coefficients. That two dimensional sequence is entered to the pattern matching stage of the speech recognition system where it is classified using a given set of patterns or models and an Euclidean-type measure of similarity.

In recent years, speech researchers have found out that the discrimination capacity of the classifier can be strongly enhanced by properly processing the two dimensional spectral representation. This is done in the quefrencial dimension by means of cepstral liftering (Juang et al., 1987), and in the temporal dimension by means of the so-called dynamic features or differential parameters (Furui, 1986).

On the one hand, cepstral liftering actually involves filtering the log spectrum in the sense of performing a periodic convolution of it with the Fourier transform of the cepstral window or lifter. On the other hand, each differential parameter can be envisioned as the output of a linear filter driven by the time sequence of a cepstral coefficient (or any other spectral parameter) where each sample corresponds to a frame. Thus, both types of processing can be interpreted from a filtering viewpoint in either frequency or time, and the frequencial analysis of that filtering operation permits to gain an useful insight into their performance (Nadeu and Juang, 1994). In this sense, these parameters can be referred to as filtered parameters.

Every filter used so far in the frequency dimension (or, analogously, in the time dimension) shows band-pass characteristics. Furthermore, it has two basic components: (1) a differentiation component, that corresponds to a kind of high-pass liftering (or filtering), and (2) a smoothing component that performs a low-pass liftering (or filtering). The differentiation component produces an augmentation of the frequency (or time) resolution – in the sense of a dynamic amplification – of the spectrum. The smoothing component of the filter attenuates the likely unreliable high quefrency (or frequency) components. Consequently, we can interpret that filtering as changing the tradeoff between frequency (or time) resolution and error power of the spectral estimation process involved in the parameterization, in order to enhance the discrimination capability of the speech recognizer.

The order of the LPC model is another parameter that allows to control the same kind of tradeoff, since an increased order may yield a higher frequency resolution of the speech representation, but may also produce an augmentation of the spectral estimation error. When the speech signal is noisy, the tradeoff may depend on its SNR and the noise characteristics.

Three different cepstral lifters are considered in our experiments:

$$\begin{aligned}
&\text{Bandpass:} && w(n) = 1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right);\\
&\text{Slope:} && w(n) = n;\\
&\text{Inverse of standard deviation:} && w(n) = \frac{1}{\sigma_{c(n)}};
\end{aligned} \tag{6}$$

where $n = 1,\dots,L$, and $\sigma_{c(n)}$ is the standard deviation of the nth cepstral coefficient $c(n)$. If p denotes the prediction filter order, the value of L is typically $3p/2$ for the bandpass lifter (Juang et al., 1987), and p for both the slope lifter (Hanson and Wakita, 1987) and the inverse of the standard deviation lifter (Tohkura, 1987).

Probably the most common version of time filtered parameter is the so-called regression coefficient or delta-cepstrum (Furui, 1986). Its associated impulse response is the first degree discrete Legendre polynomial, i.e.,

$$h(n) = \begin{cases} -n, & -N \le n \le N, \\ 0, & \text{elsewhere}. \end{cases} \tag{7}$$

In our work, we applied this filter to the time sequences of cepstral coefficients to obtain for each frame a vector of filtered parameters that supplements the cepstral vector. The filtered energy (delta-energy) was also considered in some experiments. The length $2N+1$ of its impulse response was varied to empirically optimize it.

3.2. Multilabeling

In the Discrete Hidden Markov Model (DHMM) approach, the conventional VQ technique is applied. For each incoming vector the quantizer performs a hard decision about which of its codewords is the best match, and so the information about how the incoming vector matches other codewords is discarded (Tseng et al., 1987). When the speech is corrupted by noise, the vector of parameters can be displaced in such a way that the best match is achieved with a different codeword from that one of clean speech. The random character of that displacement is a potential source of misrecognition.

Unlike the conventional VQ, multilabeling makes a soft decision about which codeword is closest to the input vector, generating an output vector whose components indicate the relative closeness of each codeword to the input.

Let the codewords of the multilabeling codebook be $\{\boldsymbol{v}_k\}_{k=1,\dots,C}$, where C is the codebook size, and let the liftered cepstral vector in the instant t be $\boldsymbol{x}_t$. The multilabeling codebook used in this work (Hernando et al., 1993) maps the input vector $\boldsymbol{x}_t$ into an output vector $\boldsymbol{O}_t = \{w(\boldsymbol{x}_t,\boldsymbol{v}_k)\}_{k=1,\dots,C}$, whose components are calculated with

$$w(\boldsymbol{x}_t,\boldsymbol{v}_k) = \frac{1/d(\boldsymbol{x}_t,\boldsymbol{v}_k)}{\sum_{m=1}^{C} 1/d(\boldsymbol{x}_t,\boldsymbol{v}_m)}, \tag{8}$$

where $d(\boldsymbol{x}_t,\boldsymbol{v}_k)$ is the distance between $\boldsymbol{v}_k$ and $\boldsymbol{x}_t$. These components are positive, their sum is 1 and they decrease with $d(\boldsymbol{x}_t,\boldsymbol{v}_k)$. Thus, they provide a heuristic measure describing the likelihood that the input vector $\boldsymbol{x}_t$ would be derived from the class represented by the codeword $\boldsymbol{v}_k$. Under the standard DHMM approach, $w(\boldsymbol{x}_t,\boldsymbol{v}_k)$ would take value 1 for the codeword with the best match and value 0 for the rest.

The DHMM algorithms must be generalized to accommodate this multilabeling output. For a given state j of the HMM, the probability that a vector $\boldsymbol{x}_t$ is observed can be written as

$$b_j(\boldsymbol{x}_t) = \sum_{k=1}^{C} w(\boldsymbol{x}_t,\boldsymbol{v}_k)\, b_j(k), \tag{9}$$

where $b_j(k)$ denotes the discrete output probability associated with the codeword $\boldsymbol{v}_k$ and the state j.

Forward–backward and Viterbi algorithms are simply generalized using Eq. (9) instead of $b_j(k)$. With respect to the training problem, Baum–Welch reestimation formulas for the transition probabilities a_{ij} and initial state probabilities π_i are generalized in the same manner.

Regarding the reestimation of $b_j(k)$, the maximum likelihood estimation leads to the following formula for a training sequence of length T:

$$b_j'(k) = \frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j) \frac{w(\boldsymbol{x}_t,\boldsymbol{v}_k)b_j(k)}{\sum_{k=1}^{C} w(\boldsymbol{x}_t,\boldsymbol{v}_k)b_j(k)}}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}, \tag{10}$$

where $\alpha_t(j)$ and $\beta_t(j)$ are, respectively, the well known forward and backward probabilities, and T is the utterance length. In Eq. (10) $b_j'(k)$ receives the contribution of the probability $\alpha_t(j)\beta_t(j)$ (probability of being in the state j at the time t) weighted by

$$\frac{w(\boldsymbol{x}_t,\boldsymbol{v}_k)b_j(k)}{\sum_{k=1}^{C} w(\boldsymbol{x}_t,\boldsymbol{v}_k)b_j(k)}. \tag{11}$$

This weight can be interpreted as the normalized contribution of the codeword $\boldsymbol{v}_k$ to the observation probability $b_j(\boldsymbol{x}_t)$.

Nevertheless, better speech recognition scores were obtained just by using the reestimation formula (Hernando, 1993):

$$b_j''(k) = \frac{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)w(\boldsymbol{x}_t,\boldsymbol{v}_k)}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)}. \tag{12}$$

Under this new formulation, the probability $\alpha_t(j)\beta_t(j)$ contributes to $b_j''(k)$ according to the heuristic likelihood $w(\boldsymbol{x}_t,\boldsymbol{v}_k)$. Actually, this is the reestimation formula employed in the experiments reported in this work.

The application of Eq. (12) leads to output probability distributions smoother than ones corresponding to the application of Eq. (10). The output probabilities of DHMM present an intermediate degree of smoothing. In fact, Eq. (12) can be interpreted as a distance-based smoothing technique (Sugawara et al., 1985) of discrete output probabilities, where the window $w(\boldsymbol{x}_t,\boldsymbol{v}_k)$ has been adapted to each codeword.

Although Eq. (12) does not guarantee the convergence of the training process, in practice its use decreases the required number of iterations because Eq. (12) reduces the dependence on previous values of the output probabilities $b_j(k)$. Furthermore, Eqs. (8) and (9) can be simplified using only the K most significant values of $w(\boldsymbol{x}_t,\boldsymbol{v}_k)$ for each $\boldsymbol{x}_t$, where K is lower than the codebook size C. The corresponding reductions in computational load make the MultiLabeling Hidden Markov Model (MLHMM) approach extremely efficient.

The presented multilabeling method is similar to those described in (Tseng et al., 1987) and (Nishimura and Toshioka, 1987). The main discrepancies with respect to them are the possibility of using any distance measure between cepstral vectors – i.e. Euclidean or projection – in Eq. (8), the alternative generalization of HMM algorithms (Eq. (12)), and the simplification of Eqs. (8) and (9) by using only the K closest codewords.

The SemiContinuous HMM approach (SCHMM) (Huang, 1992) also is closely related to the multilabeling approach. However, the components of the output vector $w(\boldsymbol{x}_t,\boldsymbol{v}_k)$, that are estimated from an heuristic viewpoint in the multilabeling method, are estimated from a stochastic viewpoint in the SCHMM. Concretely, whereas in the multilabeling approach the codewords are the centroid (mean) of each cluster, in the semicontinuous approach, the codebook is modeled as a parametric family of mixture Gaussian densities, characterized by the mean and the variance of each cluster.

The recognition rates obtained with both MLHMM and SCHMM approaches are similar and outperform considerably those obtained using DHMM, both in clean conditions and in the presence of additive white noise (Hernando et al., 1993).

3.3. Cepstral projection distance

If it is known that the reference and test signals have different degrees of noise corruption, there is no obvious reason to maintain the symmetry characteristics of the conventional Euclidean distance on the liftered cepstral vectors, commonly used in speech recognition.

Analytical derivations and empirical observations performed by Mansour and Juang (1989b) 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.

Using a speech recognition system based on the Dynamic Time Warping (DTW) approach, the best results (Mansour and Juang, 1989b) were obtained using

$$d_{\mathrm{P}} = |\boldsymbol{C}_{\mathrm{t}}|(1 - \cos\beta) = |\boldsymbol{C}_{\mathrm{t}}| - \frac{\boldsymbol{C}_{\mathrm{t}}^{\mathrm{T}}\boldsymbol{C}_{\mathrm{r}}}{|\boldsymbol{C}_{\mathrm{r}}|}, \tag{13}$$

where $\boldsymbol{C}_{\mathrm{t}}$ and $\boldsymbol{C}_{\mathrm{r}}$ are the liftered cepstral column vectors of the test and reference templates being compared in one step of the DTW algorithm, β is the angle between them, $|\cdot|$ denotes norm and $^{\mathrm{T}}$ denotes transposition.

This projection distance is used in some experimental results reported in Section 4. In these experiments, an HMM-VQ recognition system is used. In such a system, when the projection distance of Eq. (13) is chosen, a problem arises during the codebook training process since a closed formula has not been derived to compute the centroid that minimizes the global distortion of a cluster. In our tests, we used as centroid the mean of the distribution (that minimizes the global distortion in the case of Euclidean distance) since it yielded good results in preliminary recognition experiments with additive white noise (Hernando and Nadeu, 1991).

4. Experimental results

This section reports the experimental results obtained by applying the OSALPC representation and all the similarity measuring techniques reviewed in Section 3, both separately and in combination, to recognize isolated words in a real noisy car environment, in a multispeaker task, using an HMM-VQ recognition system and training in clean conditions (Hernando and Nadeu, 1994).

4.1. Database and recognition system

The database used in the 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, seated in the passenger's seat. The signals were recorded through a back-electret microphone that was centered on the passenger's sunvisor in different noisy conditions: 5 repetitions with the engine and the fan off and 20 more with the engine on and different fan positions, of which 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 average SNR values for each of the different conditions were 12, 15 and 2 dB for 0 (car stopped, engine on), 70 and 130 km/h, respectively. In all conditions, the car noise is not flat (Lecomte et al., 1989), but it does not show any periodicities: it is broad-band noise. The system was trained with the signals uttered when the engine and the fan were off, i.e., in noise free conditions, and the noisy signals were used for testing.

The analog speech signal was sampled at 8 kHz and 12 bits quantized. Then the digital signal was manually endpointed and preemphasized with $1 - 0.95\,z^{-1}$. In the parameterization stage, the signal was divided into frames of 30 ms at a rate of 15 ms and each frame was characterized by L liftered cepstral parameters, obtained either by the conventional LPC method or the new OSALPC technique. In some tests, the time filtered parameters of the frame were also obtained. Each information was separately vector-quantized using a codebook of 64 codewords by means of either conventional VQ or the multilabeling method, with either the conventional Euclidean distance or the new cepstral projection distortion measure. Each digit was characterized by a first-order, left-to-right, hidden 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 in empirically optimizing the model order and the type of cepstral lifter using the cepstral Euclidean distance upon the cepstral vectors and conventional VQ. Preliminary recognition results in noise free conditions showed that neither the model order nor the type of cepstral lifter are relevant for our task.

However, in the presence of noise the recognition results are very sensitive to the model order and the type of cepstral lifter. The results can bee seen in Table 1, for conventional LPC, and in Table 2, for the novel OSALPC representation. In both cases, the results are shown in terms of the car speed, for model order p equal to 8, 12 and 16, and for bandpass, slope and inverse of standard deviation (ISD) lifters.

The best results were obtained using prediction order equals to 16 and either the ISD lifter for the conventional LPC or the slope lifter for the OSALPC technique, i.e. a relatively high prediction order and a non-symmetrical cepstral lifter for both kinds of parameterization. It is worth noting that, although the average SNR value when the car is running at 70 km/h is a little higher than SNR with the car stopped (engine on), the recognition rates are worse when the car is running. This may be explained by the different articulatory effects due to the environment.

In Fig. 4, recognition rates obtained using these optimum orders and lifters are compared, in terms of car speed, with those obtained using prediction order equal to 8 and bandpass lifter. Notice that the results are very sensitive to those factors, and that relatively high prediction orders and non-symmetrical cepstral lifters are clearly preferable in noisy conditions. It can also be seen that, using the optimum orders and lifters, OSALPC noticeably outperforms LPC in severe noisy conditions.

Table 1
Recognition rates using LPC-cepstrum for several prediction orders and lifters

Order	Lift./speed	0	70	130
8	Bandpass	93.7	88.9	58.2
	Slope	93.2	85.1	59.7
	ISD	93.0	84.5	61.2
12	Bandpass	96.7	93.9	71.0
	Slope	93.2	91.4	75.2
	ISD	95.2	84.1	60.0
16	Bandpass	90.7	86.1	66.2
	Slope	92.7	85.6	72.0
	ISD	97.5	92.1	79.0

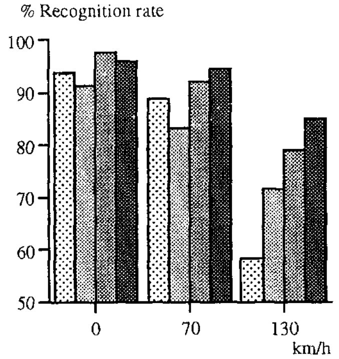
Table 2
Recognition rates using OSALPC-cepstrum for several prediction orders and lifters

Order	Lift./speed	0	70	130
8	Bandpass	91.2	83.4	71.7
	Slope	91.7	85.1	68.0
	ISD	93.5	82.6	72.2
12	Bandpass	96.7	89.3	74.5
	Slope	91.0	87.1	76.2
	ISD	95.5	87.9	77.2
16	Bandpass	92.7	85.5	69.7
	Slope	96.0	94.6	85.0
	ISD	95.5	91.2	80.7

Regarding the time filtered parameters, the use of delta-cepstrum (Δc) and delta-energy (ΔE), in the case of the conventional LPC parameterization, and the use of delta-cepstrum, in the case of the OSALPC technique, provided excellent results. The best results were obtained using a window length of 240 ms for the estimation of filtered parameters.

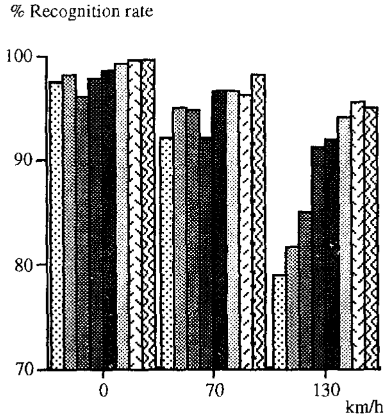
On the other hand, the recognition rates obtained with both MLHMM and SCHMM approaches were similar and outperformed considerably those obtained using DHMM. In particular, MLHMM approach using the reestimation Eq. (12) led to slight better results than SCHMM ones, with lower computational load. Because of this fact, that approach was used in final recognition experiments.

It is worth noting that in both MLHMM and SCHMM approaches, the parameters of the codebook and the models can be jointly optimized to achieve an optimal model/codebook combination. When this mutual optimization of models and codebook is made, the various cepstral lifters have almost no effects on the SCHMM performance since SCHMM adjust to any kind of liftering by modifying the variance estimates. However, this mutual optimization has not been considered in this work because of its computational complexity. In this case, the various cepstral lifters affect the performance of both MLHMM and SCHMM approaches.



Techniques/Speed	0	70	130
░ LPC-8, bandpass	93.7	98.9	58.2
▒ OSALPC-8, bandp.	91.2	83.4	71.7
▓ LPC-16, ISD	97.5	92.1	79.0
█ OSALPC-16, slope	96.0	94.6	85.0

Fig. 4. Optimization of prediction order and cepstral liftering in LPC and OSALPC techniques.



Techniques/Speed	0	70	130
░ LPC-D	97.5	92.1	79.0
▒ LPC-ML	98.2	94.9	81.7
▓ OSALPC-D	96.0	94.7	85.0
▇ OSALPC-ML	97.7	92.1	91.2
■ LPC-D-Δ	98.5	96.6	92.0
□ LPC-ML-Δ	99.2	96.6	94.0
▨ OSALPC-D-Δ	99.5	96.1	95.5
▩ OSALPC-ML-Δ	99.5	98.1	95.0

Fig. 5. Comparison of several combinations of techniques.

Finally, the results obtained using the cepstral projection distance were not better than those obtained applying the Euclidean distance. When the optimum prediction order and cepstral lifter for each parameterization technique were used, the recognition rates using the projection distance were 95.2, 88.9 and 67.2% – for speeds of 0, 70 and 130 km/h, respectively –, in the case of conventional LPC, and 95.5, 93.3 and 77.2%, in the case of the OSALPC representation.

The combination of all these techniques, except the cepstral projection distance measure, provided even better results than those obtained applying each technique separately. In Fig. 5, recognition rates obtained with optimum orders and lifters are compared in terms of the employed kinds of parameterization – LPC or OSALPC – and vector quantization – conventional VQ in discrete (D) HMM or multilabeling (ML) –, with or without time filtered parameters (Δ). The various combinations of techniques have been ordered taking into account the recognition rates obtained in severe noisy conditions.

As can be observed in Fig. 5, when no delta-parameters are used the OSALPC technique obtains excellent results in severe noisy conditions, but the conventional LPC technique results are better than OSALPC results when the car is stopped. However, using delta-cepstrum, OSALPC outperforms LPC in all the considered conditions. The best results are obtained using OSALPC parameterization, delta-cepstrum and multilabeling.

5. Conclusions

From the application of the OSALPC parameterization in combination with several robust similarity measuring techniques to speech recognition in a real noisy car environment with an HMM-VQ system, some conclusions can be summarized.

(a) The cepstral representation based on linear prediction of the one-sided autocorrelation sequence (OSALPC) has shown to be attractive because of both its high recognition performance with respect to the conventional LPC in severe noise conditions, and its computational simplicity. This technique relies on the fact that the autocorrelation sequence is less affected by broad-band noise – as in the case of noisy car environment – than the signal itself, especially at high lag indices. In the frequency domain, the spectrum of the one-sided autocorrelation sequence enhances the highest power frequency bands, and thus it is more robust to broad-band noise than speech spectrum itself.

(b) When linear prediction techniques are used, the use of a non-symmetrical lifter and a relatively high prediction order is preferable. Actually, a non-symmetrical cepstral lifter is more convenient in the presence of broad-band noise due to the fact that cepstral coefficients of lower index are more affected by this type of noise than higher order ones in the truncated cepstral vector. On the other hand, 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 the fact that the sensitivity to this type of noise tends to decrease along the autocorrelation lag. Too high model orders, however, yield poor recognition results because of the presence of spurious peaks in the spectral estimates. Regarding the time filtered parameters, that are routinely used in current speech recognition systems, their inclusion is very useful in all the considered conditions.

(c) The multilabeling technique clearly outperforms the conventional VQ method. Unlike the conventional VQ, multilabeling makes a soft decision about which codeword is closest to the input vector, generating an output vector whose components indicate the relative closeness of each codeword to the input. This information is especially important in the case of noisy signals, because the hard decision performed by the conventional VQ about which of the codewords is the best match may easily be affected by the noise. The recognition rates obtained with both MultiLabeling models (MLHMM) described in this paper and the closely related SemiContinuous models (SCHMM) were similar and outperformed considerably those obtained using discrete models (DHMM). In particular, the algorithm proposed in this work to estimate the MLHMM parameters tends to smooth the output probability distributions and led to slight better results than SCHMM models in this task, with lower computational load.

(d) The cepstral projection distance measure, that was proposed for the case of additive white noise, does not show a good performance in this particular noisy car environment.

(e) Finally, the combination of the various techniques, except the cepstral projection measure, provides better results than those obtained applying each technique separately. The best results are obtained using OSALPC parameterization, delta-cepstrum and multilabeling.

Acknowledgements

The authors would like to thank students J. Cobo and J. Dachs for helping in computer tasks, and G. Babini for kindly allowing us to use the Italian ESPRIT-ARS database.

References

P. Alexandre and P. Lockwood (1993), "Root cepstral analysis: A unified view. Application to speech processing in car noise environments", *Speech Communication*, Vol. 12, No. 3, pp. 277–288.

J.A. Cadzow (1982), "Spectral estimation: An overdetermined rational model equation approach", *Proc. IEEE*, Vol. 70, pp. 907–939.

S. Furui (1986), "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 34, pp. 52–59.

B.A. Hanson and H. Wakita (1987), "Spectral slope distance measures with linear prediction analysis for word recognition in noise", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 35, pp. 968–973.

J. Hernando (1993), Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos, Ph.D. Dissertation, Dept. Signal Theory and Communications, Polytechnical University of Catalonia, Barcelona.

J. Hernando and C. Nadeu (1991), "A comparative study of parameters and distances for noisy speech recognition", *Proc. Eurospeech'91*, September 1991, Genoa, pp. 91–94.

J. Hernando and C. Nadeu (1994), "Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques", *Proc. Internat. Conf. Acoust. Speech Signal Process.'94*, Adelaide, April 1994, Vol. II, pp. 69–72.

J. Hernando, C. Nadeu and E. Lleida (1992), "On the AR modelling of the one-sided autocorrelation sequence for noisy speech recognition", *Proc. ICSLP'92*, Banff, October 1992, pp. 1593–1596.

J. Hernando, J.B. Mariño and C. Nadeu (1993), "Multiple multilabeling to improve HMM-based speech recognition in noise", *Proc. Eurospeech'93*, Berlin, September 1993, pp. 1643–1646.

X.D. Huang (1992), "Phoneme classification using semicontinuous hidden Markov models", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 40, pp. 1062–1067.

F. Itakura (1975), "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 23, pp. 67–72.

B.H. Juang (1991), "Speech recognition in adverse conditions", *Computer Speech and Language*, Vol. 5, pp. 275–294.

B.H. Juang, L.R. Rabiner and J.G. Wilpon (1987), "On the use of band-pass liftering in speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 35, pp. 947–954.

M.A. Lagunas and M. Amengual (1987), "Non-linear spectral estimation", *Proc. Internat. Conf. Acoust. Speech Signal Process.'87*, Dallas, TX, April 1987, pp. 2035–2038.

I. Lecomte et al. (1989), "Car noise processing for speech input", *Proc. Internat. Conf. Acoust. Speech Signal Process.'89*, Glasgow, May 1989, pp. 512–515.

D. Mansour and B.H. Juang (1989a), "The short-time modified coherence representation and its application for noisy speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 37, pp. 795–804.

D. Mansour and B.H. Juang (1989b), "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 37, pp. 1959–1971.

S.L. Marple, Jr. (1987), *Digital Spectral Analysis with Applications* (Prentice-Hall, Englewood Cliffs, NJ, 1987).

H. Matsumoto and H. Imai (1986), "Comparative study of various spectrum matching measures on noise robustness", *Proc. Internat. Conf. Acoust. Speech Signal Process.'86*, Tokyo, April 1986, pp. 769–772.

D.P. McGinn and D.H. Johnson (1983), "Reduction of all-pole parameter estimation bias by successive autocorrelation", *Proc. Internat. Conf. Acoust. Speech Signal Process.'83*, Boston, MA, April 1983, pp. 1088–1091.

C. Nadeu and B.H. Juang (1994), "Filtering of spectral parameters for speech recognition", *Proc. ICSLP'94*, Yokohama, September 1994, pp. 1927–30.

M. Nishimura and K. Toshioka (1987), "HMM-based speech recognition using multi-dimensional multi-labeling", *Proc. Internat. Conf. Acoust. Speech Signal Process.'87*, Dallas, TX, April 1987, pp. 1163–1166.

K. Sugawara, M. Nishimura, K. Toshioka, M. Okochi and T. Kaneko (1985), "Isolated word recognition using hidden Markov models", *Proc. Internat. Conf. Acoust. Speech Signal Process.'85*, Tampa, FL, March 1985, pp. 1–4.

H.P. Tseng, M.J. Sabin and E.A. Lee (1987), "Fuzzy vector quantization applied to hidden Markov modeling", *Proc. Internat. Conf. Acoust. Speech Signal Process.-87*, Dallas, TX, April 1987, pp. 641–644.

Y. Tohkura (1987), "A weighted cepstral distance measure for speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 35, pp. 1414–1422.