SPEAKER IDENTIFICATION IN NOISY CONDITIONS USING LINEAR PREDICTION OF THE ONE-SIDED AUTOCORRELATION SEQUENCE*

Javier Hernando, Climent Nadeu, Carlos Villagrasa and Enric Monte

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

ABSTRACT

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 noisy conditions. In this paper the OSALPC technique is applied to the problem of speaker identification in noisy conditions. As shown with experimental results, using additive white noise, that technique also achieves much better results than both LPC and mel-cepstrum parameterizations in this task.

1. INTRODUCTION

The performance of existing speech and speaker recognition systems degrades rapidly in the presence of background noise when training and testing cannot be done in the same ambient conditions. In order to develop a 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 [1]. However, speech and speaker recognition in noisy environments remains an unsolved problem.

One of the main attempts to combat the noise problem consists of finding novel acoustic representations that are resistant to noise corruption in order to replace the traditional parameterization techniques, which are known to be very sensitive to the presence of noise.

Concretely, in speech and speaker recognition applications, cepstral-based parameters are the most common representations. However, the most widely used techniques of computing cepstra, such as linear prediction (LPC) [2] and mel-cepstrum [3], lead to poor recognition rates in noisy conditions, even if only a modest level of noise contamination is present in the speech signal.

Recently, the authors proposed an alternative parameterization technique called One-Sided Autocorrelation Linear Predictive Coding (OSALPC) [4] for noisy speech recognition. This technique, closely related with the Short-Time Modified Coherence (SMC) representation [5], is essentially an AR modeling of the causal part of the autocorrelation sequence and its use in noisy speech recognition has shown to be attractive because of its simplicity and high speech recognition performance with respect to the standard LPC in severe conditions of additive white noise [4] and noisy car environment [6].

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

In this paper the OSALPC technique is applied to the problem of text independent speaker identification in noisy conditions in order to gain some perspective of the merit of that technique with respect to the conventional LPC and mel-cepstrum parameterization techniques. Experiments have been carried out using a simple and efficient speaker identification system that uses an arithmetic-harmonic sphericity measure on covariance matrices [7]. In this work, only additive white noise has been considered.

The paper is organized in the following way. In section 2 the OSALPC technique is revised and its relationship with the standard LPC approach and the SMC representation is discused. Section 3 is dedicated to report the experiments and results. Finally, in section 4 some conclusions are summarized from those results.

2. OSALPC REPRESENTATION

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

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

which verifies

$$R^+(m) + R^+(-m) = R(m), \quad -\infty \le m \le \infty \qquad (2)$$

Its Fourier transform is the complex spectrum

$$S^+(\omega) = \frac{1}{2}[S(\omega) + jS_H(\omega)] \qquad (3)$$

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

$$E(\omega) = |S^+(\omega)| \qquad (4)$$

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 components lying outside the enhanced frequency band 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 [9].

These both properties, robustness to noise and pole-preservation, suggest 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, proposed in [4] as a robust representation of speech signal when noise is present.

The algorithm to calculate the cepstrum coefficients corresponding to the OSALPC technique is simple:

a) Firstly, from the speech frame of length N the autocorrelation lags until M = N/2 are computed.

b) In the case of additive white noise, as in this paper, R(0) is set to 0 because it is very corrupted by noise.

c) Secondly, the Hamming window from m=0 to M is applied on the one-sided autocorrelation sequence obtained in steps a) and b).

d) Thirdly, the first p + 1 autocorrelation lags of this sequence are computed from m = 0 to p using the classical biased estimator.

e) Then these values are used as entries to the Levinson-Durbin algorithm to estimate the AR parameters.

f) Finally, the cepstral coeffcients corresponding to the model are recurrently computed from those AR parameters.

A block diagram of the proposed algorithm is given in Fig. 1.

The robustness of that algorithm to additive white noise is illustrated in Fig. 2. In that case, the conventional biased autocorrelation estimator, i.e. the one that is commonly employed in speech processing, was used to compute the one-sided autocorrelation sequence. As it can be seen in the figure, the OSALPC square envelope strongly enhances the highest power frequency band and is more robust to additive white noise than the LPC spectrum.

It can also be seen in Fig. 2 that spurious peaks appear in the OSALPC square envelope. Probably, they are due to the fact that OSALPC technique does not actually perform a deconvolution between the filter and the excitation of the canonic model of speech production done by the standard LPC technique [4].

However, in spite of the OSALPC technique only performs a partial deconvolution, its use in noisy speech recognition has shown to be attractive because of its simplicity and high speech recognition performance with respect to the conventional LPC in severe conditions of additive white noise [4] and noisy car environment [6].

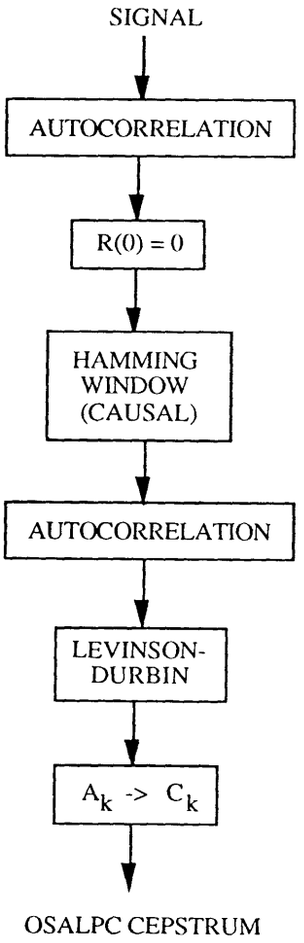
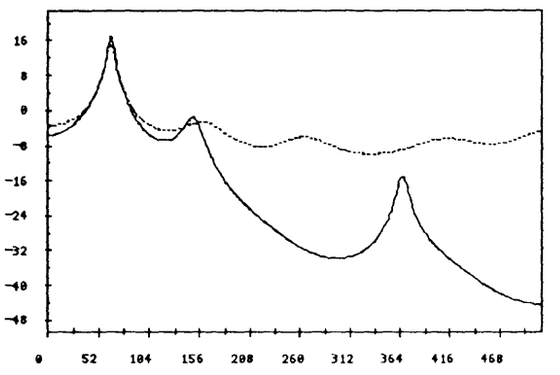
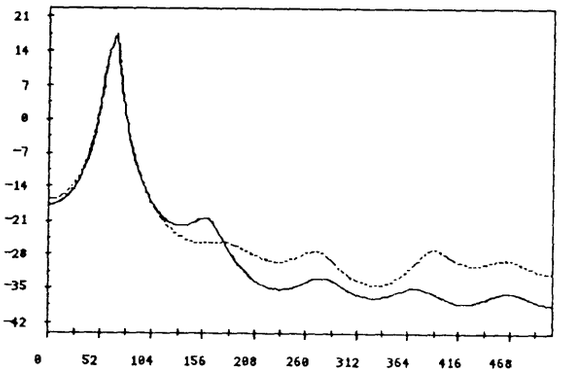


Fig.1. Block diagram for the calculation of the OSALPC cepstrum (R(0) is set to zero only in the case of additive white noise)



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 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 [5], 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 [4] on noisy speech recognition with both the conventional 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 and in a recent work on speech recognition in real noisy car environment [6], OSALPC was implemented using the same one-sided autocorrelation estimator than SMC (i.e., the coherence estimator, which is defined in [5]), 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 [10].

3. SPEAKER IDENTIFICATION EXPERIMENTS

This section reports the application of the OSALPC technique to the problem of text independent speaker identification in noisy conditions, i.e. from conversational speech utterances that have been corrupted by noise, in order to gain some perspective of the merit of that technique with respect to the standard parameterization techniques, LPC and mel-cepstrum.

3.1. Speech Database

For training text independent speaker identification, a task-specific corpus is not necessary. The testing corpus is different from the training corpus. Consequently, the speaker identification task becomes more difficult, but it is closer to a real application.

The TIMIT [11] database was used in our experiments. It consists of 420 speakers (130 females and 290 males) classified into the eight "dialect regions" of American English. Each speaker utters 10 sentences. 2 of these sentences are "dialect sentences" and are uttered by every speaker. The 8 others are different for each speaker: 5 "MIT" sentences (from a set of 450 sentences), designed to provide a rich variety of phonetic segments and contexts; and 3 "TI" sentences (from 1890), taken from a large corpus of written text. There are no impostors, only cooperative speakers. In the average, a sentence lasts about 3 seconds. The speech signal is sampled at 16 KHz and quantized using 2 bytes per sample.

The TIMIT sentences were considered clean signals. Noisy speech was simulated by adding zero mean white Gaussian noise to the clean signals.

3.2. Speaker Identification System

There are a number of techniques that have demonstrated good text independent speaker identification performance in relatively low-noise environments. In this paper, the experiments have been carried out using a simple speaker identification system that uses an arithmetic-harmonic sphericity measure on the covariance matrices of the sequence of the parameter vectors [7], which is easy to implement and computationally efficient.

In that system, one reference is used per speaker, which is the covariance matrix of the acoustic parameters of a training utterance.

The arithmetic-harmonic sphericity distance measure between a test covariance matrix $\mathbf{Y}$ and a reference covariance matrix $\mathbf{X}$ is defined as:

$$\mu(\mathbf{X},\mathbf{Y}) = \log\left(\frac{\mathbf{A}}{\mathbf{H}}\right) \qquad (5)$$

where $\mathbf{A}$ and $\mathbf{H}$ are respectively the arithmetic and harmonic means of the eigenvalues of $\mathbf{Y}$ relative to $\mathbf{X}$ (eigenvalues of the product $\mathbf{YX}^{-1}$), that are always possitive. This measure is non-negative and equals to zero iff $\mathbf{A} = \mathbf{H}$, that is iff all eigenvalues are equal (i.e. when $\mathbf{X}$ and $\mathbf{Y}$ are proportional). Moreover, μ is clearly symmetric. Another important property of this measure comes from the fact that it can be computed very efficiently without an explicit computation of the eigenvalues of $\mathbf{Y}$ relative to $\mathbf{X}$.

That measure is used in association with the 1-nearest neighbor decision rule. The possibility of rejection is not taken into account.

3.3. Experiments and Results

For each speaker, the reference covariance matrix was computed from the concatenated acoustic parameters corresponding to the 5 "MIT" sentences of the TIMIT database, recorded in noise-free conditions. Thus, a training utterance lasts approximately 5 x 3 = 15 seconds.

During testing two experiments were performed per speaker. For each experiment, the test covariance matrix was computed from the concatenated parameters of 2 of the 5 remaining sentences. Therefore, a test utterance lasts approximately 2 x 3 = 6 seconds. Those signals were contaminated by zero mean white Gaussian noise in order that the SNR becomes ∞ (clean), 30, 20 and 10 dB.

The speech signal was preemphasized with $1 - 0.95\,z^{-1}$. In the parameterization stage of the system, the signal was divided into frames of 25 ms at 10 ms rate. Each frame was characterized by 20 cepstral parameters calculated by either the conventional LPC method, the mel-cepstrum technique or the OSALPC representation proposed in this paper. In all cases, an analysis order (prediction order in the LPC and OSALPC methods and number of filters in the mel-cepstrum technique) equal to 20 was used.

Table 1 shows the speaker identification scores for a set of 100 speakers (200 tests) corresponding to the three parameterizations above considered -LPCC (Linear Prediction Cepstral Coefficients), MFCC (Mel-Frequency

SNR (dB)	clean	30	20	10
LPCC	100	95.0	55.0	7.0
MFCC	100	95.5	53.0	19.5
OSALPCC	100	98.5	79.0	20.5

Table 1. Speaker identification scores for 100 speakers

Cepstral Coefficients) and OSALPCC (One-Sided Autocorrelation Linear Prediction Cepstral Coefficients)- in terms of the SNR of additive white noise of the test signals. The reference signals are clean.

As it can be seen in Table 1, the speaker identification results are very sensitive to noise distortion. The same observation was made in [12]. Relating to the parameterization technique, the speaker identification scores obtained using LPCC and MFCC are very similar, but OSALPCC results are much better whenever the SNR is not too low. Experiments with other sizes of speaker sets showed similar relative scores between the three techniques.

In Fig. 3, the results of Table 1 are compared with the results corresponding to the conventional LPC technique in the case that both test and training data, although degraded by noise, are matched in terms of SNR. This case is called "reference" in the figure because, although it is a somewhat artificial situation, provides a "good target for any noise compensation or adaptation scheme" [12].

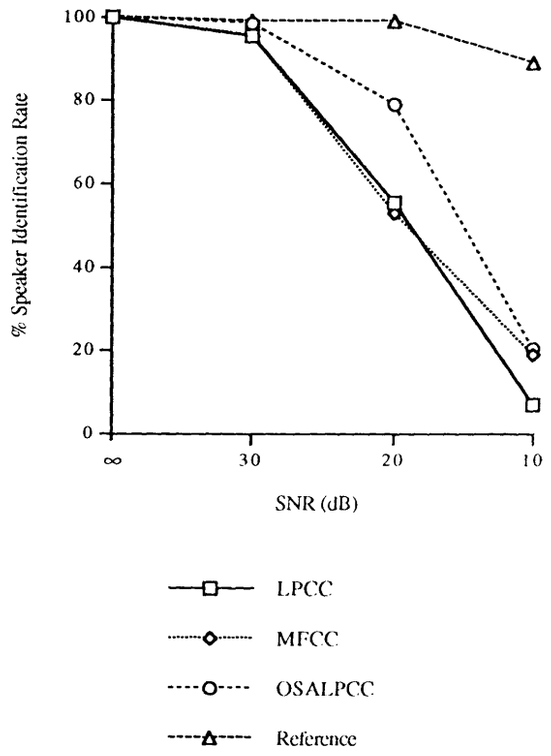


Fig. 3. Comparison of techniques

In the experiments of Table 1, every technique achieved 100 % speaker identification rate in clean conditions. In order to gain some perspective of the merit of the OSALPC technique in these conditions, experiments were performed for a set of 200 speakers (400 tests) using the conventional LPC and the new OSALPC techniques. The speaker identification scores obtained were 99.75 and 99.5, respectively. This lost of speaker identification accuracy of the OSALPC technique in clean conditions is probably due to the imperfect deconvolution of the speech signal performed by that technique.

4. CONCLUSIONS

In this paper, the linear prediction of the one-sided autocorrelation sequence (OSALPC), already proposed by the authors in [4] for noisy speech recognition, is applied to text independent speaker identification in the presence of additive white noise and it is compared with the traditional parameterizations of speech, LPC-cepstrum and mel-cepstrum. Experiments have been carried out using a simple and efficient speaker identification system that uses an arithmetic-harmonic sphericity measure on covariance matrices [7].

From this study, two main conclusions are attained:

a) Text independent speaker identification using the LPC-cepstrum and mel-cepstrum techniques, degrades drastically in the presence of additive white noise.

b) The OSALPC technique noticeably outperforms those techniques whenever the SNR is not too low.

Therefore, we can assert that the OSALPC technique is also useful, at least with this system, for speaker identification in noisy conditions.

REFERENCES

[1] B.H. Juang, "Speech Recognition in Adverse Conditions", Computer Speech and Language, Vo. 5, pp. 275-294, 1991.

[2] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. on ASSP, Vo. 23, pp. 67-72, Feb. 1975.

[3] S.B. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. ASSP, Vo. 28, pp. 357-366, 1980.

[4] J. Hernando, C. Nadeu, E. Lleida, "On the AR Modelling of the One-Sided Autocorrelation Sequence for Noisy Speech Recognition", Proc. ICSLP'92, Banff (Canada), pp. 1593-1596, Oct. 1992.

[5] D. Mansour and B.H. Juang, "The Short-Time Modified Coherence Representation and its Application for Noisy Speech Recognition", IEEE Trans. on ASSP, Vo. 37, pp. 795-804, Jun. 1989.

[6] J. Hernando, C. Nadeu, "Speech Recognition in Noisy Car Environment Based on OSALPC Representation and Robust Similarity Measuring Techniques", Proc. ICASSP'94, Adelaida (Australia), Apr. 1994. pp. II-69-72.

[7] F. Bimbot, L. Mathan, "Text-Free Speaker Recognition Using an Arithmetic-Harmonic Sphericity Measure", Proc. EUROSPEECH'93, Belin (Germany), Sept. 1993, pp. 169-172.

[8] M.A. Lagunas and M. Amengual, "Non-Linear Spectral Estimation", Proc. ICASSP'87, Dallas, pp. 2035-38, Apr. 1987.

[9] D.P. McGinn and D.H. Johnson, "Reduction of All-Pole Parameter Estimation Bias by Successive Autocorrelation", Proc. ICASSP'83, Boston, pp.1088-91, Apr. 1983.

[10] J. Hernando, "Técnicas de Procesado y Representación de la Señal de Voz para el Reconocimiento del Habla en Ambientes Ruidosos", Ph.D. Dissertation, Dpt. Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona. May 1993.

[11] W. Fisher, V. Zue, J. Bernstein, D. Pallet, "An Acoustic-Phonetic Data Base", J. Acoust. Soc. Amer. Suppl. (A), 81, S92, 1986.

[12] J.P. Openshaw, Z.P. Sun, J.S. Mason, "A Comparison of Composite Features under Degraded Speech in Speaker Recognition", Proc. ICASSP'93, Minneapolis, pp. II-371-374, Abril 1993.