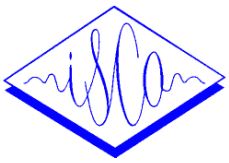


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

# ON THE AR MODELLING OF THE ONE-SIDED AUTOCORRELATION SEQUENCE FOR NOISY SPEECH RECOGNITION

Javier Hernando, Climent Nadeu and Eduardo Lleida

Dept. of Signal Theory and Communications
Polytechnical University of Catalonia
08034 Barcelona, Catalonia, Spain

## ABSTRACT

Speech recognition in noisy environments remains an unsolved problem even in the case of isolated word recognition with small vocabularies. Recently, several techniques have been proposed to alleviate this problem. Concretely, two closely related parameterization techniques based on an AR modelling in the autocorrelation domain called SMC [1] and OSALPC [2] have shown good results using speech contaminated by additive white noise. The aim of this paper is twofold: to compare several techniques based on an AR modelling in the autocorrelation domain, including SMC and OSALPC, and to find the optimum model order and cepstral liftering for noisy conditions.

## 1. INTRODUCTION

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

A spectral estimation technique widely used in speech processing and, particularly, in speech recognition is linear predictive coding (LPC) [3], equivalent to an AR modelling of the signal. Concretely, recent contributions [4] have showed that the use of a bandpass liftering of the LPC-cepstral coefficients in the standard Euclidean distance measure can lead to excellent results in noise free conditions. However, the standard LPC technique is known to be very sensitive to the presence of additive noise. This fact yields poor recognition rates in noisy conditions when these techniques are applied.

For recognition in noisy speech, Hanson and Wakita [4] applied to LPC-spectra the spectral slope distance measure, which shows high correlation with subjective phonetic distinctions and is equivalent to a quefrency weighting on the cepstral domain. This slope lifter is concerned with the fact that, in the presence of white or broad-band noise, lower order cepstral coefficients are more affected than higher order terms in the truncated cepstral vector.

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

Recently, Mansour and Juang have proposed [1] the SMC (Short-Time Modified Coherence) technique for robust spectral analysis of speech, based on the well known fact that the autocorrelation sequence is less affected by noise than the original signal. The SMC representation is essentially an AR modelling in the autocorrelation domain and outperforms considerably the standard LPC approach in speech recognition in severe noisy conditions.

In [2] the authors presented a parameterization technique called OSALPC (One-Sided Autocorrelation Linear Predictive Coding) as a robust representation of speech signals when noise is present. This technique, closely related with the SMC representation and with the use of an overdetermined set of Yule-Walker equations proposed by Cadzow in [6] to seek rational models of time series, is interesting in noisy speech recognition because of its simplicity, computational efficiency and high recognition accuracy.

This 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 other parameterizations based on an AR modelling in the autocorrelation domain is discussed. Section 3 reports the application of all the liftering and parameterization techniques mentioned to an isolated word multispeaker recognition task using the HMM approach in order to compare their performance in the presence of additive white noise. Finally, in section 4 some conclusions are summarized.

## 2. AR MODELLING IN THE AUTOCORRELATION DOMAIN

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" $E(\omega)$ [7] can be defined as

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

This envelope characteristic, along with the high dynamic range of speech spectra, originate that $E(\omega)$ strongly enhances the highest power frequency bands. Thus, the noise components lying outside the enhanced frequency band are largely attenuated in $E(\omega)$ with respect to $S(\omega)$ (see Fig.1). On the other hand, it is well known that $R^+(n)$ has the same poles than the signal [8].

It is then suggested that the AR parameters of the signal can be more reliably estimated from $R^+(n)$ than directly from the signal itself when it is corrupted by noise. In the same manner as LPC performs a linear prediction of the speech signal, we may consider a linear prediction of $R^+(n)$. This is the basis of OSALPC (One-Sided Autocorrelation Linear Predictive Coding) parameterization technique proposed in [2] as a robust representation of speech signal when noise is present.

Before describing the algorithm of this technique, let us explore now the implications of applying linear prediction on the causal part of the autocorrelation sequence. Firstly, let us assume that the speech signal x(n), whose autocorrelation is R(n), is given by the linear convolution

10.21437/ICSLP.1992-396

$$x(n) = h(n) * e(n) \qquad (5)$$

where h(n) is the impulse response of a pth-order all-pole filter driven by e(n), and e(n) is assumed to be a train of impulses for voiced sounds and white noise for unvoiced sounds. If

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \qquad (6)$$

is the z-transform of h(n) and $S_e(\omega)$ is the power spectrum of e(n), it follows that

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + e(n) \qquad (7)$$

$$S(\omega) = \frac{S_e(\omega)}{|A(\omega)|^2} \qquad (8)$$

The standard LPC approach performs a deconvolution of the speech signal since, assuming that $S_e(\omega)$ is a constant in (8), it obtains the characteristics of the vocal tract filter, H(z). In this case, e(n) is the linear prediction error of the signal x(n).

As $R^+(n)$ has the same poles than the signal, if $B(\omega)$ is the Fourier transform of the driving function that obtains $R^+(n)$ at the output of the filter H(z), we can write

$$S^+(\omega) = \frac{B(\omega)}{A(\omega)} \qquad E^2(\omega) = \frac{|B(\omega)|^2}{|A(\omega)|^2} \qquad (9)$$

Thus, the OSALPC representation is equivalent to assume that $B(\omega)$ is constant in (9) and performs an AR modelling of the square envelope $E^2(\omega)$.

Let us explore now the meaning of the above assumption. From (3) we can write $S(\omega)$ as a function of $A(\omega)$ and $B(\omega)$ as follows

$$S(\omega) = S^+(\omega) + (S^+(\omega))^* = \frac{B(\omega)}{A(\omega)} + \frac{B^*(\omega)}{A^*(\omega)} \qquad (10)$$

and from identification of (10) and (8) it results that

$$S_e(\omega) = B(\omega)\, A^*(\omega) + B^*(\omega)\, A(\omega) \qquad (11)$$

i.e., $B(\omega)$ depends on both $S_e(\omega)$ and $A(\omega)$ and can no longer be considered a constant. Thus, we can assert that OSALPC technique does not actually perform a deconvolution between filter and excitation as does the LPC of the speech signal [9]. However, in spite of the OSALPC technique only performs a partial deconvolution, as it will be seen its use in speech recognition outperforms the standard LPC approach for noisy speech.

For the calculation of the OSALPC representation it has been implemented a simple and efficient algorithm:

a) Firstly, from the speech frame of length N the autocorrelation lags from m =1 to M = N/2 are calculated using the classical biased autocorrelation estimator, generally used in speech processing. In the case of additive white noise, as in this paper, R(0) is set to 0 because it is very corrupted by noise.

b) Secondly, the Hamming window is applied on the one-sided autocorrelation sequence obtained in the first step.

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

d) Finally, these values are used as entries to the Levinson-Durbin algorithm to estimate the AR parameters.

In order to compare this technique with other related techniques based on an AR modelling in the autocorrelation domain proposed for robust spectral estimation, it is convenient to notice that the Levinson-Durbin algorithm found the AR parameters that minimize the mean square value of the linear prediction error of the windowed sequence whose autocorrelations lags -computed using the classical biased estimator- are the entries of the algorithm.

So the AR parameters obtained in the OSALPC representation will minimize the norm of the error vector of the matrix equation (12), in which R(n) denotes the windowed value of $R^+(n)$ in order to simplify the notation (notice that the autocorrelation sequence and the one-side autocorrelation sequence are identical for the range of values used in the equation) and $\varepsilon(n)$ denotes the linear prediction error of this sequence.

$$\begin{pmatrix} R(1) & 0 & 0 & \dots & 0 \\ R(2) & R(1) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p) & R(p-1) & R(p-2) & \dots & 0 \\ R(p+1) & R(p) & R(p-1) & \dots & R(1) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ R(M) & R(M-1) & R(M-2) & \dots & R(M-p) \\ 0 & R(M) & R(M-1) & \dots & R(M-p+1) \\ 0 & 0 & R(M) & \dots & R(M-p+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & R(M) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(p) \\ \varepsilon(p+1) \\ \vdots \\ \vdots \\ \varepsilon(M) \\ \varepsilon(M+1) \\ \varepsilon(M+2) \\ \vdots \\ \varepsilon(M+p) \end{pmatrix} \qquad (12)$$

Using this formulation, it is easy to compare the OSALPC technique with the use of an overdetermined set of Yule-Walker equations, proposed by Cadzow in [6] to seek rational models, and the SMC representation, proposed recently by Mansour and Juang [1] for robust spectral analysis of speech

For an AR process x(n), it is well known that its autocorrelation sequence R(n) obeys the following difference expression

$$R(m) = -\sum_{k=1}^{p} a_k R(m-k) \qquad (13)$$

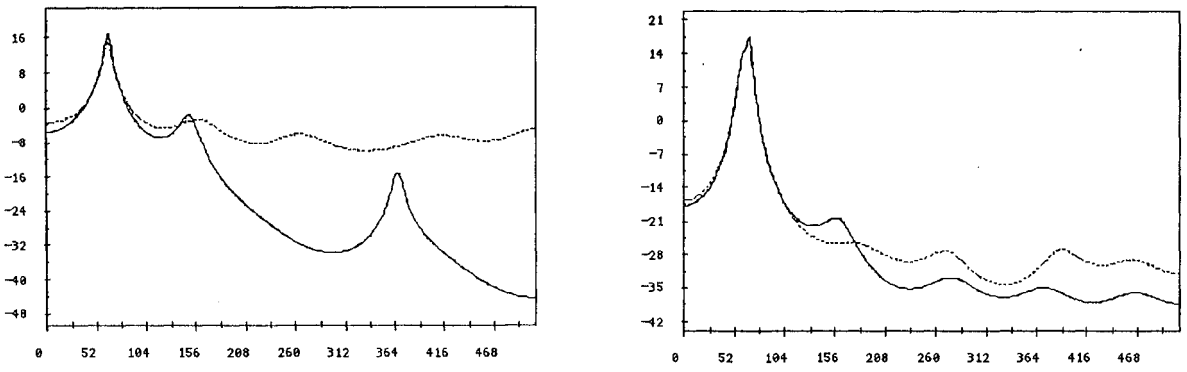
for m > 0.

As is well known the resolution of the first p equations that provides this expression, for m = 1 to p, is the basis of the standard LPC approach, using the classical biased autocorrelation estimator on the windowed signal. This determined set of equations is known as Yule-Walker equations (YWE).

Cadzow proposed the use of more than the minimal number of equations of (13) forming an overdetermined set of Yule-Walker equations (ref. in this paper as OYWE) to reduce the "undesired parameter hypersensitivity" [6]. The AR parameters obtained with this method will minimize the norm of the error vector in the following matrix equation, where M is the number of equations:

$$\begin{pmatrix} R(1) & R(0) & R(1) & \dots & R(p-1) \\ R(2) & R(1) & R(0) & \dots & R(p-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p) & R(p-1) & R(p-2) & \dots & R(0) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ R(M) & R(M-1) & R(M-2) & \dots & R(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(p) \\ \vdots \\ \varepsilon(M) \end{pmatrix} \qquad (14)$$

On the other hand, also it is well known that for an AR process x(n) contaminated by additive white noise its autocorrelation sequence R(n) only obeys (13) for m > p. The first p equations, for m = p+1 to 2p are known as the High Order Yule-Walker equations (HOYWE) and it is possible to apply in this case the same idea as




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 noisy free conditions (solid line) and SNR = 0 dB (dotted line).

above and arrive to an overdetermined set of HOYWE (ref. in this paper as OHOYWE)

$$\begin{pmatrix} R(p+1) & R(p) & R(p-1) & \dots & R(1) \\ R(p+2) & R(p+1) & R(p) & \dots & R(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(2p) & R(2p-1) & R(2p-2) & \dots & R(p) \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ R(M) & R(M-1) & R(M-2) & \dots & R(M-p) \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \varepsilon(p+1) \\ \varepsilon(p+2) \\ \vdots \\ \varepsilon(2p) \\ \vdots \\ \vdots \\ \varepsilon(M) \end{pmatrix} \quad (15)$$

From the matrix equations (12), (14) and (15) it is clear the relationship between OSALPC, OYWE and OHOYWE representations. In the three techniques a linear prediction is performed on an autocorrelation sequence. The only main difference between them is the range of autocorrelation lags considered in the minimization of the prediction error, apart from details of implementation that have not been relevant in the preliminary recognition experiments carried out (e.g., in the OYWE and OHOYWE techniques the autocorrelation sequence is not usually windowed).

On the other hand, the SMC technique is also based on an AR modelling in the autocorrelation domain and there are only two algorithmic differences between this technique and OSALPC. Firstly, the SMC representation uses a covariance estimator instead of the classical biased estimator to compute the first autocorrelation sequence. Secondly, the autocorrelation entries to the Levinson-Durbin algorithm in the SMC representation are calculated in the frequency domain using a spectral shaper in the form of a square root. In terms of the above OSALPC formulation, that difference actually consists of an AR modelling of the envelope $E(\omega)$ instead of $E^2(\omega)$.

In spite of the similarity between all these techniques, as it will be seen in next section, the OSALPC representation outperforms considerably the OYWE, OHOYWE and SMC techniques in speech recognition in severe noisy conditions. On the other hand, with respect to the computational efficiency of the algorithms, OSALPC and SMC techniques are much more efficient than OYWE and OHOYWE techniques because they make use of the Levinson-Durbin algorithm.

## 3. SPEECH RECOGNITION EXPERIMENTS

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

### 3.1. Speech database and recognition system

The database used in our experiments consists of ten repetitions of the Catalan digits uttered by seven male and three female speakers (1000 words) and recorded in a quiet room. Firstly, the system was trained with half of the database and tested with the other half. Then the roles of both halves were changed and the reported results were obtained by averaging the two results.

The analog speech was first bandpass filtered to 100-3400 Hz. by an antialiasing filter, sampled at 8 KHz and quantized using two bytes per sample. The digitized clean speech was manually endpointed to determine the boundaries of each word. The endpoints obtained in this way were used in all our experiments including those in which noise was added to the signal. In this way we eliminate the effect of errors in endpoint detection on recognition accuracy and focus only on the recognition process itself. Clean speech was used for training in all the experiments. Noisy speech was simulated by adding zero mean white Gaussian noise to the clean signal so that the SNR of the resulting signal becomes $\infty$ (clean), 20, 10 and 0 dB. No preemphasis was performed.

In the parameterization stage of the recognition system, the signal was divided into frames of 30 ms. at a rate of 15 ms. and each frame was characterized by L cepstral parameters obtained either by the standard LPC method or the other techniques exposed in last section. Before entering the recognition stage, the cepstral parameters were vector-quantized by means a codebook of 64 codewords using the standard Euclidean distance measure between liftered cepstral vectors. This codebook size had been optimized in preliminary experiments using the standard LPC technique.

Each digit is characterized by a first order, left-to-right, discrete Markov model. The tradeoff between computational load and recognition accuracy led us to consider models of 10 states without skips.

### 3.2. Recognition results

The first experiments carried out with the above described speech recognition system consisted of empirically optimizing the model order and the type of cepstral lifter in the standard LPC technique. The preliminary recognition results showed that neither the model 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. In table I, the recognition results for LPC model order p = 8, 12 and 16 and rectangular, bandpass and slope lifters are presented.

Table I. Recognition rates for LPC technique

| SNR (dB) | | ∞ | 20 | 10 | 0 |
|---|---|---|---|---|---|
| p=8 | Rectang. | 99.8 | 74.2 | 36.6 | 22.2 |
| | Bandpass | 99.8 | 92.8 | 56.8 | 27.0 |
| | Slope | 99.7 | 95.7 | 72.3 | 34.1 |
| p=12 | Rectang. | 99.8 | 66.1 | 34.0 | 22.8 |
| | Bandpass | 99.7 | 96.2 | 73.7 | 29.0 |
| | Slope | 99.8 | 98.9 | 89.5 | 54.2 |
| p=16 | Rectang. | 99.9 | 73.0 | 35.5 | 22.2 |
| | Bandpass | 100 | 94.0 | 60.2 | 19.6 |
| | Slope | 99.8 | 93.2 | 70.7 | 41.2 |

It is clear from the table that the slope lifter outperforms the rectangular and bandpass lifters for every model order. It is concerned with the fact that in the presence of white noise lower order cepstral coefficients are more affected than higher order terms in the truncated cepstral vector. On the other hand, using the slope lifter recognition rates were calculated for a big range of values of the model order and the best results were those obtained for p = 12. The convenience of this high relatively high order is due to the fact that lower order autocorrelation lags are more affected by additive white noise than higher order lags. Model orders too high, however, yield poor recognitions results because of the appearance of spurious peaks in the spectral estimation.

In table II, the recognition rates of all the LPC-based parameterization techniques mentioned in this paper are presented, using the same value of M (= N/2) and the optimum model order and lifter for the standard LPC technique, i.e., p = 12 and slope lifter. Obviously, these are not the optimum conditions for each parameterization technique but the results can help to compare their performance. Moreover, in preliminary experiments it was found that the OYWE, OHOYWE, SMC and OSALPC techniques are less sensitive to changes in the model order and the type of cepstral lifter than the standard LPC approach.

Table II. Recognition rates for several LPC-based techniques. p=12 and slope lifter

| SNR (dB) | ∞ | 20 | 10 | 0 |
|---|---|---|---|---|
| LPC | 99.8 | 98.9 | 89.5 | 54.2 |
| OYWE | 99.9 | 95.9 | 66.9 | 31.7 |
| OHOYWE | 99.5 | 97.7 | 81.3 | 43.1 |
| SMC | 99.0 | 97.0 | 89.2 | 67.5 |
| OSALPC | 98.6 | 97.7 | 93.7 | 75.9 |

It is clear from the table that the recognition rates of the OSALPC and SMC representations are excellent and outperform considerably the other techniques in severe noisy conditions. However, in noise free conditions there is a lost of recognition accuracy due to the imperfect deconvolution of the the speech signal performed by these techniques. On the other hand, the OSALPC representation outperforms the SMC technique in noisy conditions in spite of the major simplicity of the OSALPC technique.

## 4. CONCLUSIONS

In this paper a comparison of several LPC-based techniques in the autocorrelation domain is made for noisy speech recognition. The optimum model order and cepstral liftering in noisy conditions also has been investigated. From this study, two main conclusions are attained:

a) using the standard LPC approach, a relatively high model order and the slope cepstral lifter are preferable in noisy conditions; and

b) the OSALPC technique, based on the LPC autocorrelation method applied on the one-sided autocorrelation sequence, yields the best results among all the compared LPC-based techniques in severe noisy conditions and is less sensitive than the standard LPC approach to changes in the model order and the type of cepstral lifter.

## ACKNOWLEDGEMENTS

The authors would like to thank Jordi Cobo and David Riu for their help in the software development.

## REFERENCES

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

[2] J. Hernando and C. Nadeu, "A comparative study of parameters and distances for noisy speech recognition", EUROSPEECH'91, Genova, September 1991, pp. 91-94.

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

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

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

[6] J.A. Cadzow, "Spectral estimation: an overdetermined rational model equation approach", Proc. of IEEE, vol.70, Sept. 1982, pp. 907-939.

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

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

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

This work was supported by the TIC grant number 92-0800-C05-04