AR modelling of the speech autocorrelation to improve noisy speech recognition

J. Hernando and C. Nadeu

Department of Signal Theory and Communications. Polytechnical University of Catalonia
Ap. 30002, 08080 Barcelona, Spain.

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

A spectral estimation technique widely used in speech processing and, particularly, in speech recognition is linear predictive coding (LPC), equivalent to an AR modelling of the signal. Concretely, recent contributions [3] 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 is equivalent to a slope liftering. As well known, 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) parameterization for noisy speech recognition, based on the well known fact that the autocorrelation sequence is less affected by noise than the original signal. This technique is essentially an AR modelling in the autocorrelation domain.

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 [5] to seek rational models of time series, is interesting in noisy speech recognition because of its simplicity, computational efficiency and high recognition accuracy. In this paper, OSALPC is revised and its relationship with all these techniques is discussed. Also, their performance in noisy speech recognition is compared.

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

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

which verifies

$$R^+(m) + R^+(-m) = R(m), \quad -8 = m = 8 \quad (2)$$

This paper has been supported by the grant TIC 92-0800-C05/04

Its Fourier transform is the complex spectrum

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

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" $E(\omega)$ 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.

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. This is the basis of OSALPC (One-Sided Autocorrelation Linear Predictive Coding) technique proposed in [2] as a robust representation of noisy speech signals.

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

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

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 $H(z)$.

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 the Fourier transform of $R^+(n)$ and its spectrum as

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

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 [6]. 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.

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

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.

Let us now compare OSALPC with other related techniques based on an AR modelling in the autocorrelation domain: the SMC representation, proposed recently by Mansour and Juang [1] for robust spectral analysis of speech, and the use of an overdetermined set of Yule-Walker equations, proposed by Cadzow in [5] to seek rational models.

With respect to SMC, 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 $E^2(\omega)$.

On the other hand, the relationship between OSALPC and the spectral approach proposed by Cadzow in [5] is also very close. As well known, for an AR process x(n) its autocorrelation sequence R(n) obeys for m > 0 the following difference expression

$$R(m) = -\sum_{k=1}^{p} a_k R(m-k) \qquad (12)$$

The resolution of the first p equations that this expression provides, 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 (12) forming an overdetermined set of Yule-Walker equations, for m = 1 to M, (ref. in this paper as OYWE) to reduce the "undesired parameter hypersensitivity" [5].

Also it is well known that for an AR process x(n) contaminated by additive white noise its autocorrelation sequence R(n) only obeys (12) 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 above and arrive to an overdetermined set of HOYWE, for m = p+1 to M (ref. in this paper as OHOYWE).

It is clear the relationship among 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.

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. RECOGNITION EXPERIMENTS

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 and sampled at 8 KHz. The digitized clean speech was manually endpointed to determine the boundaries of each word. 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 (clean), 20, 10 and 0 dB. No preemphasis was performed.

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. Each digit was characterized by a left-to-right discrete Markov model 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)		8	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,. the convenience of a relatively high model order, 12, 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)	8	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 OSALPC are excellent and outperform considerably the results of 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 this technique.

REFERENCES

1. D. Mansour and B.H. Juang, IEEE Trans. on ASSP-37, nº 6, Jun. 1989, pp. 795-804.
2. J. Hernando and C. Nadeu, EUROSPEECH´91, Genova, September 1991, pp. 91-94.
3. B.H. Juang, L.R.Rabiner and J.G. Wilpon, IEEE Trans. on ASSP-35, nº 7, Jul. 1987, pp. 947-54.
4. B.A. Hanson and H. Wakita, IEEE Trans. on ASSP-35, nº 7, Jul. 1987, pp. 968-73.
5. J.A. Cadzow, Proc. of IEEE, vol.70, Sept. 1982, pp. 907-939.
6. C. Nadeu, J. Pascual and J. Hernando, ICASSP'91, Toronto, May 1991, pp. 3677-80.