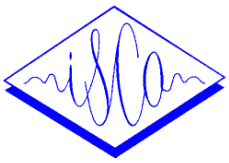

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ROBUST HOS-BASED TECHNIQUES APPLIED TO SPEECH RECOGNITION AND ENHANCEMENT

Josep M. SALAVEDRA*, Javier HERNANDO*, Enrique MASGRAU**, Asunción MORENO*

* *Department of Signal Theory and Communications. Universitat Politècnica de Catalunya. Apartat. 30002. 08080-BARCELONA. SPAIN. Tel/Fax: +34-3- 4017404 / 4016447 . E-mail: mia@tsc.upc.es*
** *Department of Electrical Engineering and Computers. Universidad de Zaragoza.*

ABSTRACT

We study some speech enhancement algorithms based on the iterative Wiener filtering method due to Lim-Oppenheim [2], where the AR spectral estimation of the speech is carried out using a second-order analysis. But in our algorithms we consider an AR estimation by means of cumulant analysis. This work extends some preceding papers due to the authors, where information of previous speech frames is taken to initiate speech AR modelling of the current frame. Two parameters are introduced to dessign Wiener filter at first iteration of this iterative algorithm. These parameters are the Interframe Factor (IF) and the Previous Frame Iteration (PFI). A detailed study of them shows they allow a very important noise suppression after processing only first iteration of this algorithm, without any appreciable increase of distortion. Finally, the simplest cumulant-based algorithm is applied to Speech Recognition and some preliminary results are presented.

1. INTRODUCTION

It is well known, that many applications of speech processing that show very high perfomance in laboratory conditions degrade dramatically when working in real environments because of low robustness. The solution we propose here concerns to a preprocessing front-end in order to enhance the speech quality by means of a speech parametric modelling insensitive to the noise. The use of HO cumulants for speech AR modelling calculation provides the desirable uncoupling between noise and speech. It is based on the property that for Gaussian processes only, all cumulants of order greater than two are identically zero [1]. Moreover, the non-Gaussian processes presenting a symmetric p.d.f. have null odd-order cumulants. Considering a Gaussian or a symmetric p.d.f. noise (a good approximation of very real environments) and the non-Gaussian characteristic of the speech (principally for the voiced frames) it would be possible to obtain an spectral AR modelling of the speech more independent of the noise by using, e.g., third-order cumulants of noisy speech instead of common second-order statistics.

2. ITERATIVE PARAMETRIC WIENER METHOD

In the original Lim-Oppenheim Method [2], noisy speech is enhanced by means of an iterative Wiener filtering. Clearly, filtered speech signal contains a smaller residual noise but it presents a larger spectral distortion. Therefore, increasing the number of iterations doesn't always involve a better speech estimation. It is well known that this algorithm leads to a narrowness and a shifting of the speech formants [3], providing an unnatural sounding speech. In [6] a detailed convergence analysis of this algorithm is carried out. It is proved that this estimated Wiener filter tends to cancel all signal frequencies with SNR lower than 4.77dB, and an additional attenuation, proporcionally to the noise level, affects signal frequencies with higher SNR, in comparison to the optimum Wiener filter. Only the non-contaminated speech frequencies undergo a null attenuation.

A parameterized Wiener filtering has been considered to have a better control over noise suppression, intelligibility loss and computational complexity, by adding two parameters ∂ and ß in the Wiener filter computation. So, we consider the following equation:

$$W_i(w) = \left(\frac{P_y}{P_y + ß.P_r}\right)^{\partial} \quad (1)$$

By varying these parameters ∂,ß, filters with different characteristics can be obtained. In [7], a detailed study of performance was reported. High values of both parameters lead to a more aggresive Wiener filter and so noise suppression is increased but distortion increases too. We found that ∂=1.0, ß=1.2 is a good trade-off among noise suppression, distortion, computational complexity and convergence speed of the iterative filtering, when third-order statistics and low SNR are considered.

AR modelling of the speech spectrum estimation is obtained from third-order cumulants. Speech AR modelling coefficients a_k are computed by solving Third-Order Yule-Walker equations [4], [5]:

$$\sum_{k=0}^{p} a_k \cdot C_3(i-k, j) = 0 \quad ,, \; 1 < i \le p \; ; \; -p \le j \le 0 \quad (2)$$

where p=10 is the order of the AR filter. This procedure, that considers p+1 cumulant slices, presents a full-rank solution and it is unique [5].

As discussed in preceding works due to the authors [7], [8], we obtain a twofold benefit by considering this third-order AR modelling: Firstly, an accelerated convergence of the iterative algorithm and so a reduction of both computational complexity and intelligibility loss; Secondly, achievement of a non polluted AR speech parameterization. In comparison to second-order statistics estimation we obtain a good improvement but the price we pay for these advantages is a higher distortion. Thus a higher "peaking" or "narrowness" effect of the speech formants is brought about [6].

In Fig.1 an uniform improvement, iteration by iteration, is obtained when classical second-order statistics algorithm is evaluated. This improvement is similar when different values of Signal-to-Noise-Ratios are simulated. We may conclude noise suppression saturates after 4 or 5 iterations of the iterative Wiener Algorithm, because other effects, such as intelligibility loss, overcome noise reduction.

This work was supported by TIC 92-0800-C05-04

While the improvement of second-order approach increases gradually, but slowly, iteration by iteration, third-order one gets a very good improvement, about 3dB, after only two iterations and thus it obtains a faster convergence. In Fig.2, a lower noise sensitivity may be observed in the Third-Order Statistics domain: saturation effect appears after only 2 or 3 iterations in low SNR environments, and noise reduction effect is over-riding just in the first iteration when medium and high SNR environments are simulated.

3. THE INTERFRAME FACTOR IF

In Fig.1, we may appreciate an improvement that increases gradually iteration by iteration. Most part of noise reduction is obtained after processing two iterations. Third-order cumulants obtain an appreciable noise suppression (about 2dB in Cepstrum distance) after first iteration (see Fig.2) and then this speech modelling is enhanced enough (Cepstrum distance decreases 3.5dB) in the second iteration because it estimates Wiener filter from a cleaner speech signal. At first iteration, speech AR modelling is computed from noisy signal without any initial information about the features of speech signal corresponding to the current frame. However, we know some information of the current speech frame by considering that vocal tract features don't vary a lot between two consecutive overlapped frames. Therefore, we propose to obtain the first iteration AR coefficients as a combination between current frame AR estimation and previous frame AR coefficients. Thus, we dessign the non-causal Wiener filter, corresponding to first iteration, as a linear combination of coefficients a_k, belonging to two consecutive frames, calculated as follows:

$$A_k(n,1) = IF \cdot a_k(n,1) + (1 - IF) \cdot a_k(n-1,PFI) \qquad (3)$$
$$0 \le k \le P ; \quad 1 \le PFI \le MAXITER ; \quad 0 \le IF \le 1$$

where **n** is the current frame, PFI is the Previous Frame Iteration that we consider to help first iteration of the current frame and IF is the Interframe Factor. We write a_k when coefficients are estimated directly from a noisy speech frame and we note capital letter A_k when coefficients are coming from a linear combination of a_k . At the beginning of every speech activity we set parameter IF=1 because information of last speech frame is not related to the current speech frame. Wiener filter dessigns corresponding to the remaining iterations of the algorithm are estimated over a cleaner speech signal coming from Wiener filtering Output of previous iteration of the same frame:

$$A_k(n,iter) = a_k(n,iter) \quad ,, \quad 2 \le iter \le MAXITER \qquad (4)$$

where **iter** is the iteration number of the current frame. We have two parameters to control this linear combination. First parameter is the Interframe Factor IF that represents the amount of current speech AR estimation $a_k(n,1)$ we put in the AR modelling $A_k(n,1)$ of the filter.

Clean speech has been processed by this system and so distortion effect corresponding to the iterative algorithm has been evaluated. To avoid an appreciable increase of distortion effect all values of parameter IF lower than 0.6 must be discarded [8]. In Fig.3, first iteration of current frame corresponding to speech signal disturbed by AWGN at SNR=0dB has been processed and some different speech AR estimations of previous frame have been evaluated (ranging PFI from 1 to 5). We may come to the conclusion that values of parameter IF ranging from 0.6 to 0.8 represent a good trade-off between distortion and noise suppression. Therefore, we may achieve an improvement of 2dB in Cepstrum distance by introducing parameter IF (PFI=3 and IF=0.7) to estimate current speech AR modelling without any noticeable increase of distortion (0.25 dB) [8]. Thus, we may obtain an improvement higher than 4 dB in Cepstrum distance after processing only first iteration of the iterative Wiener filtering.

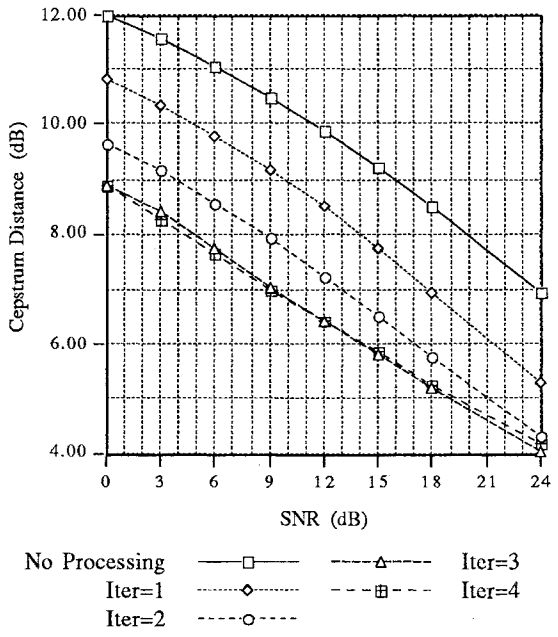

Figure 1. *Noise Suppression achieved by iterative Wiener filtering using classical autocorrelation function (AR2 Algorithm).*

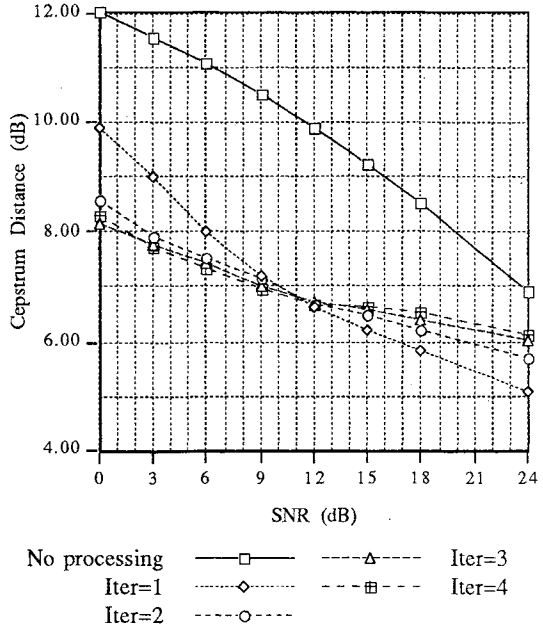

Figure 2. *Noise Suppression achieved by parameterized iterative Wiener filtering using third-order cumulants (AR3 Algorithm).*

Some different algorithms have been compared in Fig.4. We may appreciate the higher performance of AR3 algorithm in comparison to both AR2 algorithm and fourth-order cumulant-based algorithm (AR4). By introducing the Interframe Factor IF, AR3_IF algorithm has a faster convergence and so a lower computational complexity. Furthermore, in the listening tests it may be appreciated a less distortion effect. In [9], Speech AR estimation is calculated in the autocorrelation domain because One-Sided Autocorrelation function (OSA) is a pole-preserving function. Main advantage of autocorrelation domain is its lower sensitivity to background noise and therefore a very important noise reduction, after processing only 2 iterations, is achieved (about 5.5dB in Cepstrum distance). By considering parameter IF, a faster convergence and a higher noise suppression are attained. However, main disadvantage of both OSA_AR2 and OSA_AR2_IF algorithms lies in its significant distortion effect.

4. SPEECH RECOGNITION EXPERIMENTS

This section reports the application of some parameterization techniques mentioned above to recognize isolated words in a speaker-independent task, with the HMM [10] approach, in order to compare their performance in the presence of additive white noise.

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 five of the speakers and tested with the others. 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 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 10 cepstral parameters obtained either by the standard LPC method or the other techniques exposed in last section, using model order equal to 10. Obviously, these are not the optimum conditions for each parameterization technique but the results can help to compare their performance.

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 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 trade-off between computational load and recognition accuracy led us to consider models of 10 states without skips. Training and testing were performed using Baum-Welch and Viterbi algorithms, respectively [10].

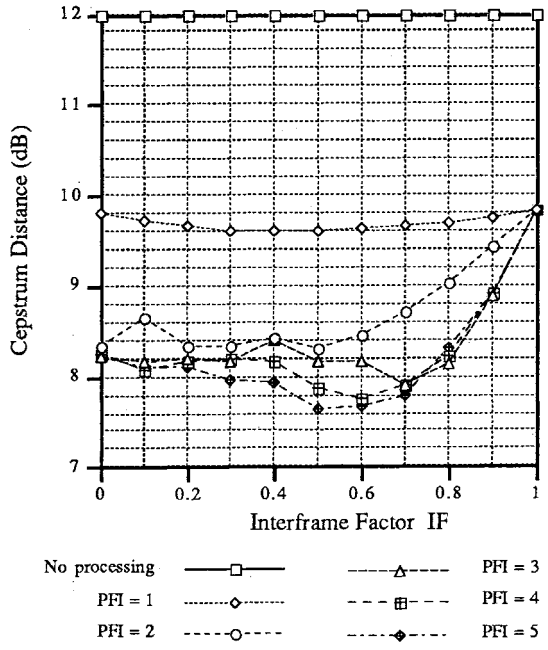

Figure 3. *Noise Suppression after processing first iteration of current frame when some different speech AR estimations belonging to different iterations of previous speech frame are considered.*

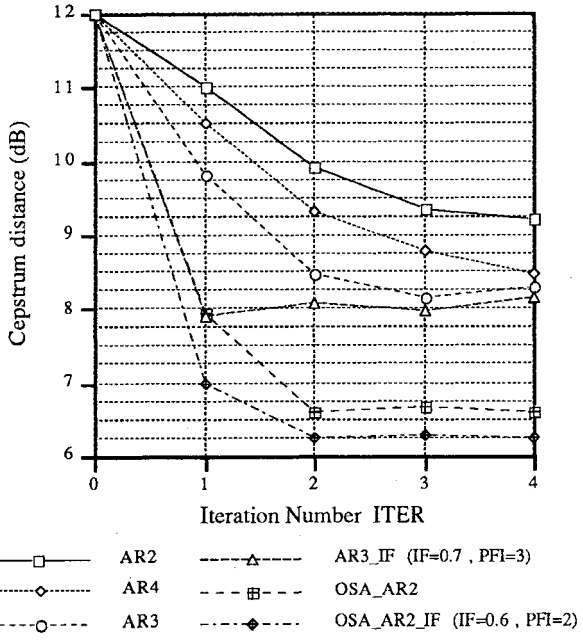

Figure 4. *Performance Comparison of different algorithms at SNR=0dB (AWGN)*

The recognition rates obtained using the standard LPC technique were 58,7 %, 37,1 % and 24 %, for 20, 10 and 0 dB of SNR, respectively. However, using the new OSALPC representation [9], based on the LPC autocorrelation method applied on the one-sided autocorrelation sequence, the corresponding results were 88,3 %, 72,6 % and 35,8 %. As it can be seen, the OSALPC results are excellent and outperform considerably standard LPC ones in all noisy conditions tested. Regarding to the other HOS-based tecniques, their recognition results are between standard LPC rates and OSALPC rates. AR3 algorithm clearly overcomes AR2 one specially in noisy environments (SNR≤10dB).

A second Test to recognize isolated words in a speaker-dependent task was performed. Firstly, the system was trained with five repetitions corresponding to all of the speakers (500 words) and it was tested with the remaining five repetitions of each speaker and every digit. We have evaluated clean speech degraded by AWG noise at four different values of SNR: 0dB, 10dB, 20dB and clean speech (∞dB). Recognition rates are presented in Table 1. These rates correspond to classical LPC technique, third-order cumulant-based AR estimation (AR3) and OSA_AR2 technique. AR3 technique achieves a recognition rate improvement of 10% at SNR=20dB in comparison to classical AR2 algorithm. It overcomes AR2 algorithm for all noise levels. Note that all of the results in Table 1 don't consider Wiener Filtering. So, we should expect a higher improvement after processing one or two iterations of the AR3 and AR3_IF algorithms (see Fig.4). OSA_AR2 algorithm outperforms the others because AR estimation is calculated in the autocorrelation domain instead of speech signal domain. A similar technique in the third-order cumulant domain is currently under study and results will be presented in future works. In short, third-order cumulants are less sensitive to noise and therefore they lead to better performance in applications belonging to Speech Recognition and Enhancement. This improvement is more important at SNR=20dB because third-order cumulants are capable to confront this noise level (see first iteration in Fig.1 and Fig.2). When noise levels are higher (SNR=0dB) we should consider an AR estimation coming from the enhanced speech signal belonging to first iteration of AR3_IF algorithm.

5. CONCLUSIONS

Some speech enhancement methods based on an iterative Wiener filtering have been proposed. Spectral estimation of speech is obtained by means of an AR modelling based on cumulant analysis to provide the desirable noise-speech uncoupling. A comparison of different order cumulant-based algorithms is given. Two parameters, IF (Interframe Factor) and PFI (Previous Frame Iteration), have been considered to take advantage of previous speech spectrum estimations to initiate AR modelling corresponding to first iteration of the current speech frame. This approach achieves a noise suppression about 4dB (Cepstrum Distance) after processing only first iteration of the AR3 algorithm. This fact represents an improvement about 2dB (Cepstrum Distance) in relation to parameterized third-order algorithm (IF=1). Finally, the convergence of the iterative algorithms based on cumulant AR estimation is strongly accelerated. Therefore, a good reduction of computational complexity and processing delay is achieved, while no appreciable increase of distortion effect is generated. All these features are specially esteemed when low and medium SNR are considered. If speech AR estimation is calculated in the autocorrelation domain, a faster convergence and a greater noise reduction are achieved. Therefore, this approach leads to the best recognition rates, but its significant distortion effect dissuades its use when enhanced speech is sent to a listener. Furthermore, the simplest cumulant-based algorithm has been integrated in a Speech Recognition System and some preliminary improvements have been reported.

SNR	0dB	10dB	20dB	∞dB
AR2	20.5	35.7	63.2	100
AR3	27.2	39.3	73.3	99.9
OSA_AR2	38.6	79.9	94.8	98.9

Table 1. *Recognition rates in a speaker-dependent context*

6. REFERENCES

[1] **C.L.Nikias, M.R.Raghuveer,** "Bispectrum Estimation: A Digital Signal Processing Framework". Proc. of IEEE, pp. 869-891. July 1987.

[2] **J.S.Lim,A.V.Oppenheim,**"All-Pole Modeling of Degraded Speech".IEEE Trans ASSP, pp.197-210. June 1978.

[3] **J.H.L.Hansen, M.A.Clements,** "Constrained Iterative Speech Enhancement with Apllications to Spech Recognition". IEEE Trans ASSP, pp.795-805. April 1991.

[4] **A.Swami, J.M.Mendel,** "AR Identifiability using Cumulants". Proc. Workshop on HO Spectral Analysis, pp.13-18. Vail, CO, USA. June 1989.

[5] **G.B.Giannakis,** "On the Identifiability of non-Gaussian ARMA Models using Cumulants". IEEE Trans ASSP, pp.1284-1296. July 1990.

[6] **E.Masgrau, J.M.Salavedra, A.Moreno, A. Ardanuy,** "Speech Enhancement by Adaptive Wiener Filtering based on Cumulant AR Modelling". Proc. ESCA Workshop on Speech Processing in Adverse Conditions, pp 143-146. Cannes, France. November 1992.

[7] **J.M.Salavedra, E.Masgrau,A.Moreno, X.Jové and J.Estarellas,** "Robust Coefficients of a Higher-order AR Modelling in a Speech Enhancement System using parameterized Wiener Filtering". Proc. MELECON'94, pp. 69-72. Antalya, Turkey. April 1994.

[8] **J.M.Salavedra, E.Masgrau, A.Moreno, J.Estarellas,** "Some robust Speech Enhancement Techniques using Higher-order AR Estimation". Proc. EUSIPCO, pp.1194-1197. Edinburgh, Scotland. September 1994.

[9] **J.M.Salavedra, E.Masgrau, A.Moreno, J.Estarellas,** "Some fast Higher-order AR Estimation Techni-ques applied to parametric Wiener Filtering". Proc. ICSLP, pp.1655-1658. Yokohama, Japan. September 1994.

[10] **L.R. Rabiner,** "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proc. IEEE, vol. 77, n.2. February 1989.