THIRD-ORDER CUMULANT-BASED WIENER FILTERING ALGORITHM APPLIED TO ROBUST SPEECH RECOGNITION

Josep M. SALAVEDRA, Javier HERNANDO

Universitat Politècnica de Catalunya. c/ Gran Capità s/n. 08034-BARCELONA. SPAIN.
Tel/Fax: +34-3-4017404 / 4016447 . E-mail: mia@gps.tsc.upc.es

ABSTRACT

In previous works [5], [6], we studied 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: a cumulant-based Wiener Filtering (AR3_IF) is applied to Robust Speech Recognition. A low complexity approach of this algorithm is tested in presence of bathroom water noise and its performance is compared to classical Spectral Subtraction method. Some results are presented when training task of the speech recognition system (HTK-MFCC) is executed under clean and noisy conditions. These results show a lower sensitivity to the presence of water noise when applying AR3_IF algorithm inside of a speech recognition task.

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.

This work was supported by CICYT under TIC95-0884-C04-02

2. ITERATIVE WIENER FILTERING

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 [4] 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. 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:

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

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.

As discussed in preceding works due to the authors [5], 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 [4].

This cumulant-based Wiener Filtering algorithm is referred as AR3_IF algorithm. Its performance, in a speech enhancement task, clearly overcomes classical Lim-Oppenheim algorithm [5], specially in low SNR environments. Next section contains some speech recognition results when this AR3_IF algorithm is implemented as a preprocessor inside of a speech recognition system.

3. RECOGNITION EXPERIMENTS

This section reports experimental results in speaker independent digit recognition. It shows an increase in recognition accuracy obtained when a noise suppression system is considered. Two different strategies of Noise Reduction Methods are compared: Spectral Subtraction (SS) and Wiener Filtering (AR3_IF) algorithms.

3.1. Database and Recognition System

The database used in the recognition experiments consists of 2 repetitions of 11 English digits ("zero", "one", "two", ..., "nine", "oh") corresponding to adult speakers (112 for training and 113 for testing) of the speaker independent digit TI [7] database, that have been recorded in clean conditions. The initial sampling frequency 20 kHz was converted to 8 kHz.

The HTK [8] recognition system, based on the Continuous-Density Hidden Markov Models (CDHMM), was appropriately prepared to perform a noise reduction task, previously to the recognition stage. In the parameterization stage, the signal was preemphasized with $1-z^{-1}$ and was divided into frames of 30ms at a rate of 10ms and each frame was characterized by 12 cepstral parameters obtained either by linear prediction (LPC), with prediction order equal to 10, or by the mel-cepstrum technique. In some tests the energy of the frame was also used. Regression analysis over 70ms was applied to the static cepstral sequence and the static energy sequence to obtain delta-cepstrum coefficients. Each digit was characterized by a first order, left-to-right, Markov model of 10 states with one mixture of diagonal covariance matrix and without skips. The same structure was used for the silence model but only with 5 states. Training was performed in two stages using Segmental k-means, with previous manual endpointing, and Baum-Welch algorithms. Testing was performed using Viterbi algorithm.

3.2. Experimental Results

In past works [6] we evaluated a speech recognition system in presence of Addtive White Gaussian Noise and we concluded that third-order cumulant-based algorithm AR3 clearly overcomes standard LPC technique, specially when noisy environments are considered (SNR≤10dB).

In this work we have evaluated the robustness of a speech recognition system in presence of water noise. This noise corresponds to typical bathroom noise when different ups and downs of a bath tap are considered. Each possibility leads to a different signal-to-noise ratio and usually low SNR are appearing.

In Table.1 we can appreciate the dramatic degradation in the test recognition rate because of the presence of water noise. Two different recognition techniques have been compared: LPCepstra and MFCC [9] techniques. Training references have been elaborated under laboratory (clean) conditions. Similar recognition rates are obtained when recognition test is executed in clean conditions. However, in presence of noise both techniques suffer a significative reduction of test recognition performance. MFCC technique seems to be more robust to the presence of noise than LPCepstra one. Therefore, this MFCC technique has been considered in all of next recognition test tasks. It must be noted that this level of noise (SNR=5dB) doesn't correspond, in fact, to the worst situation because higher levels of noise (SNR≈0dB) can appear in this environment and therefore higher performance degradations may be expected.

To avoid this sensitivity to the presence of noise in MFCC tecnique, a noise reduction system has been added at the beginning of the speech recognition system. Two different noise reduction methods have been evaluated: generalized Spectral Subtraction (SS) and Wiener Filtering based on third-order cumulant AR estimation (AR3_IF) [6].

Generalized Spectral Subtraction technique allows an estimation of the speech power spectrum P_s from the noisy speech signal x(n) and noise signal r(n) :

$$P_s(w) = \left| \max \left[P_s^{\gamma}(w) - \alpha.E\{P_r(w)\}^{\gamma} , P_0^{\gamma} \right] \right|^{1/\gamma} \quad (3)$$

where $P_0>0$ is a minimum value of P_s. Parameter γ has been set to 0.5 and two different values of parameter α have been evaluated: $\alpha=1$ (classical Spectral Subtraction) and $\alpha=2$ (overestimation of noise spectrum).

In Fig.1 some different methods have been compared: training references have been defined under clean conditions and recognition test has been processed by considering 3 different levels of water noise. MFCC technique without noise suppressor offers a recognition rate that degrades dramatically when noise level increases during recognition test. If spectral subtraction is combined with MFCC technique (referred as SS+MFCC), this degradation is reduced, specially at high levels of noise (SNR≈0dB) and recognition rate is not so bad (more than 30% better). In a similar way, MFCC combined with a low complexity Wiener Filtering AR3_IF [6] algorithm (referred as AR3_IF+MFCC) is less sensitive to changes in the level of water noise and recognition rate is about 10% better at low SNR, in comparison to previous approach.

In some real applications a noisy environment during training can be assumed. Fig.2 shows the previous technique comparison when training references have been elaborated in noisy conditions (SNR=5dB). MFCC technique without noise suppressor has a high sensitivity in front of the presence of noise. When training and test have the same level of noise, recognition performance is good (about 95% in recognition rate). But when the level of test noise changes recognition rates degrade dramatically:

SNR Train / Test	LPCepstra	MFCC
clean / clean	98.71	98.87
clean / 5dB	47.54	72.07

Table.1 : *Recognition rates in a speaker-independent context*

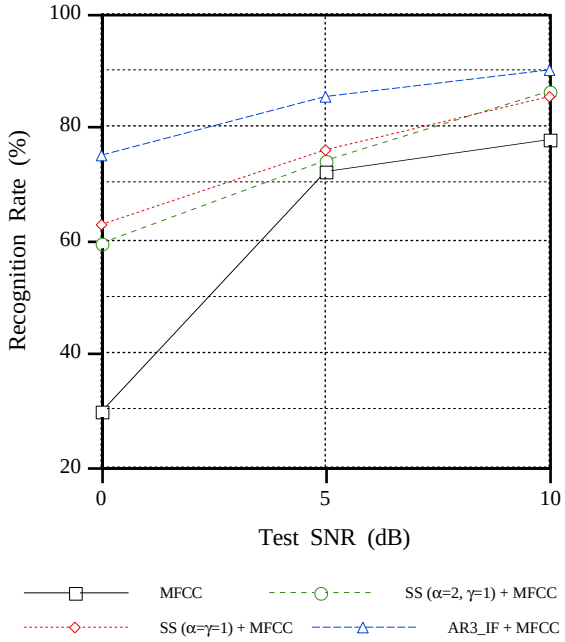

Figure.1 : *Test Recognition Rates when Training is processed with clean Speech.*

1) if the level of test noise decreases (SNR=10dB), recognition rate decreases significantly (34.7%),

2) an increase of noise level (SNR=0dB) produces a more important recognition rate degradation.

In short, MFCC technique offers a quite good performance when noise levels, during training and test, are the same. However a change in the noise level produces a more relevant degradation, in comparison to previous results shown in Fig.1 (clean conditions during training).

By adding a noise suppressor (during training and test), MFCC algorithm obtains a more similar performance in the presence of different levels of test noise. In this situation low complexity HOS-based Wiener algorithm (AR3_IF+MFCC) leads to a better recognition rate with respect to SS+MFCC technique.

4. CONCLUSIONS

A Noise Reduction Algorithm based on Wiener Filtering (AR3_IF) has been applied to a robust speech recognition task. Spectral estimation of speech in AR3_IF algorithm is obtained by means of an AR modelling based on cumulant analysis to provide the desirable noise-speech uncoupling. Recognition rate of HTK system with Mel Cepstrum coefficients degrades dramatically in noisy (bathroom) environments and therefore a noise reduction system is introduced to obtain a more robust recognition system. Recognition rates show that AR3_IF algorithm overcomes spectral subtraction (SS) approach: it enhances about 10% under clean conditions training and it increases about 1% when training is processed under noisy conditions (SNR=5dB). In short, sensitivity to the background water noise decreases significantly when HTK-MFCC recognition system contains AR3_IF algorithm as a noise reduction preprocessor.

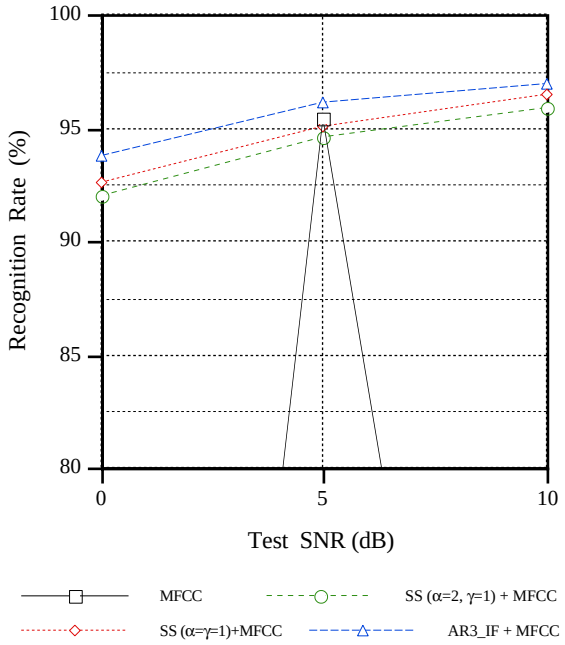

Figure.2 : *Test Recognition Rates when Training is processed with noisy Speech (SNR=5dB)*

5. 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] **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.

[5] **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.

[6] **J.M.Salavedra, J.Hernando, E.Masgrau, A.Moreno,** "Robust HOS-based Techniques applied to Speech Recognition and Enhancement". Proc. EUROSPEECH, pp.1517-1520. Madrid, Spain, September 1995.

[7] **R.G.Leonard,**"A Database for Speaker-Independent Digit Recognition". Proc. ICASSP, pp.42.11.1-4. March 1984.

[8] "**HTK**-Hidden Markov Model Toolkit v.1.5". Cambridge University Engineering Dept Speech Group and Entropic Research Labs Inc. December 1993.

[9] **S.B.Davis, P.Mermelstein,** "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". IEEE Trans. on ASSP, vol.28, No.4, pp.357-366. August 1980.