STEREOPHONIC ACOUSTIC ECHO CANCELLATION
- Application to speech recognition : some experimental results. -

F.Berthault, C.Glorion, F. Capman, J. Boudy, P. Lockwood
MATRA COMMUNICATION
Corporate R&D Division, Speech Processing Department
rue J.P. Timbaud, 78392 Bois d'Arcy Cedex, BP 26, France
email : frederic.berthault@matra-com.fr / phone : +33 1 34 60 8814

ABSTRACT

During the last decades, several adaptive filtering algorithms have been optimised in the context of speech echo cancellation. The use of such algorithms for the realisation of a hands-free function is now technically feasible for GSM mobile telephony, in car environment. Recent works now focus on the multi-channel case, and investigate the use of stereophonic echo cancellers for high-quality visioconference systems. In this paper, we address the issue of stereophonic car radio noise compensation, for speech recognition.

I. INTRODUCTION

The use of a speech recogniser in a car environment, for mobile telephony, voice-dialing, or voice-controlled functions in a car, has to face several difficulties. The speech signal is often picked-up by a hands-free microphone, located at several tens of centimeters from the mouth, resulting in a degraded signal-to-noise ratio. These problems have already been addressed in the literature for noise-robust speech recognition, [1, 7, 9]. Car noise compensation for speech recognition was previously addressed in [8]. Here, we propose to investigate and compare several stereophonic echo cancellation algorithms applied to stereophonic car-radio noise reduction for speech recognition.

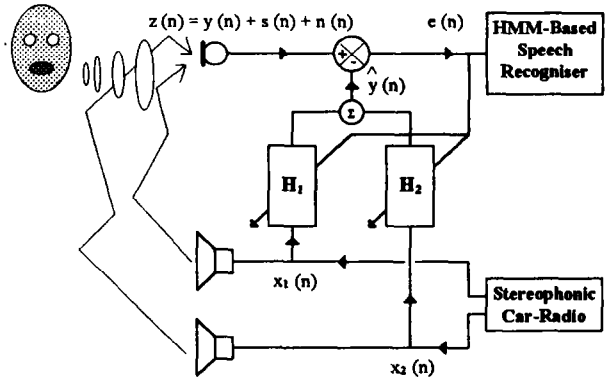

Fig I.1. : Principle of stereophonic echo cancellation for speech recognition.

The use of a stereophonic echo canceller for car radio compensation combined with a speech recognition system is described in figure (I.1). Stereophonic radio signals $x_1(n)$ and $x_2(n)$ are played through the loudspeakers. The microphone signal z(n) is composed of the corresponding stereophonic echo signal y(n), the speech signal s(n) to be recognised, and the car noise signal n(n). Each of the two echo paths is identified by an adaptive linear FIR filter, $H_1(n)$ and $H_2(n)$. The residual signal e(n) of the echo canceller is then directly used by the speech recogniser.

II. STEREOPHONIC ACOUSTIC ECHO CANCELLATION

II.1. General considerations

Theoretical studies have been carried out in several papers, [2, 10, 11], showing the different problems encountered in stereophonic echo cancellation in the context of visioconference. The cross-correlation has been identified as the main problem to deal with, in multichannel echo cancellation, [11]. Several solutions have been investigated: in [2], a two-channel algorithm taking into account cross-correlation was proposed, and in [10], a subband stereo algorithm is presented.
We propose to compare some two-channel adaptive echo cancellers in the context of speech recognition in car environment. Whereas for visioconference applications, the system is processing speech signals, a stereophonic echo canceller for car radio compensation should be able to process a wide range of signals (speech, music, ...), with very various statistical properties, and some with a high degree of non-stationarity. Such echo cancellers must also be robust to background noise related to a car environment: at 120 km/h « radio-signal »-to-noise ratio can be below 5 dB, thus reducing overall performances.

II.2. Two-channel adaptive algorithms

Four different algorithms have been considered. The first two algorithms are the well-known Normalized Least Mean Square algorithm (NLMS), and the Affine Projection Algorithm of order 2 (APA-2). The third one is an extension of the mono-channel multi-delay frequency-domain algorithm (MDFO) proposed in [4], implemented with a classical adaptation. The last one is the optimal multi-channel Fast QR-decomposition Least Square Algorithm (FQR-LS) from [3]. The FQR-LS algorithm was also combined with a Malvar-type subband structure, [12].

III. SPEECH RECOGNITION SYSTEMS

The experiments have been carried out using two types of speaker-dependent systems: an isolated-word speech recogniser and a word-spotting speech recogniser. They are based on standard continuous mixture-density left-to-right Hidden Markov Models (HMM). The extracted speech parameters are the Linear-Frequency Cepstrum Coefficients (LFCC) [7], computed every 16 ms on a 32 ms window, at a 8 kHz sampling rate. The word model is a linear sequence of 11 states, each state being represented by a single Gaussian probability density function, characterised by its mean vector and diagonal variance matrix. The feature analysis stage computes a set of 20 parameters, including LFCC cepstral coefficients, delta-cepstral coefficients and energy parameters. A speech enhancement for preprocessing speech,

based on Non-linear Spectral Subtraction (NSS) is used, as in [9].

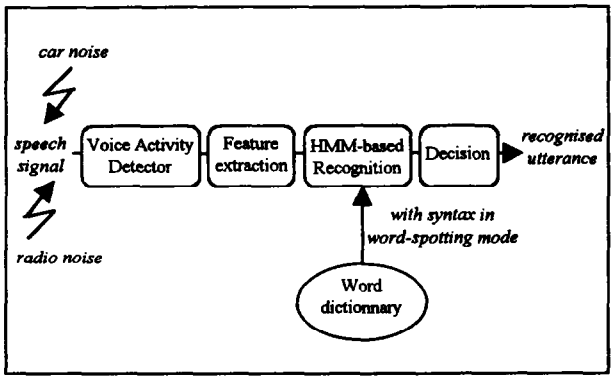

Fig III.1. : Principle of speech recognition systems.

The word-spotting speech recogniser is based on the « One-Pass » algorithm, and syntax is used during the recognition process. Noise and garbage models are used for modelling background noise and extraneous speech that may occur on either side of a valid keyword.

The training procedure is performed on the utterances pronounced in quiet conditions, and the recognition tests are performed in noisy conditions. The speech recognition principle is depicted in figure (III.1).

IV. EXPERIMENTAL RESULTS

IV.1. Databases

IV.1.1 Speech database

The speech database has 43 words (30 names and 10 command words), separated by silence pauses, and spoken by 2 female and 2 male speakers. Each sequence was repeated four times by each speaker, in a quiet environment.

IV.1.2 Radio noise database

The radio noise database consist of 29 music sequences and 11 radio sequences (news and music), recorded in in a car cockpit, in quiet conditions.

IV.1.3 Radio noise database

The car noise database includes various noise conditions (rain, open windows, ...) at several speeds (car stopped, 60, 90, 120 km/h).

IV.2. Simulations

In this paper, we present results obtained for 2 different speakers (one male, one female). 18 music sequences were used to simulate various radio noise examples (classical music, jazz, rock, pop song ...). We define by SNR1 the speech-to- « car noise » ratio, and by SNR2 the speech-to- « radio signal » ratio. The chosen values for simulated signals are for SNR1 and SNR2, 0 dB and -5dB, respectively.

The filter length of each echo channel is chosen to be 256 taps for all algorithms. The adaptation is frozen during speech sequences using theoretical labels, computed off-line and manually validated. For the algorithms comparison, cf part VI.3., theoretical labels are also given as input of the isolated-word recogniser system to avoid wrong word detection problems caused by Voice Activity Detection (VAD). This ensures a fair comparison of the adaptive algorithms.

IV.3. Echo cancellation performances

IV.3.1 Recognition rate comparison with no car noise added

Simulations performed with radio noise only showed a great improvement of recognition rates. For SNR2 = -5dB, recognition rates were improved from 45%, without radio noise compensation, to 90% when using NLMS, and over 95% when using APA, MDFO or FQR-LS algorithms.

IV.3.2 Choice of parameters used for each algorithm in car noise condition

One of the main difficulties in car radio noise compensation is the choice of the parameters used for each algorithm. Radio sequences can be so different that it is impossible to get a parameter that is optimal for all of them, as far as recognition rate is concerned.

Figures (IV.1-IV.5) show recognition rates according to parameters associated with each algorithm, for 3 different radio sequences (SNR1 = 0dB, SNR2 = -5dB).

One can see, figure (IV.1), that the optimal stepsize parameter μ of the NLMS algorithm takes different values according to the considered radio sequence. As for the APA-2 algorithm, the best recognition rates are obtained for a stepsize parameter μ between 0.1 and 0.05, whatever the sequence is, figure (IV.3), but the optimal forgetting factor λ varies according to the radio sequence, figure (IV.4). Finally,

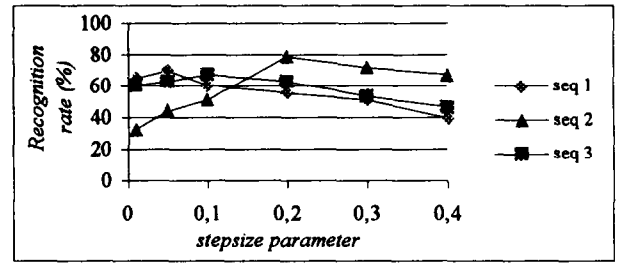

Fig IV.1 : NLMS stepsize parameter μ

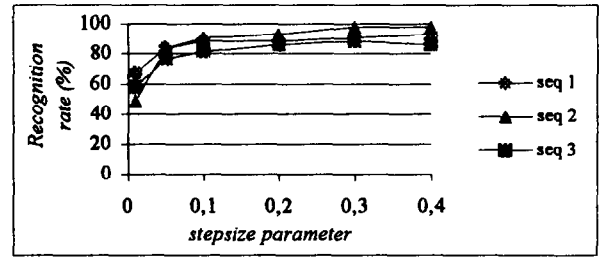

Fig IV.2 : MDFO stepsize parameter μ

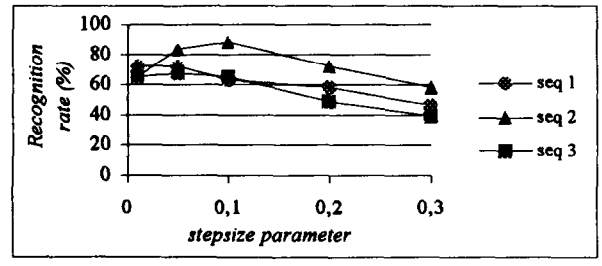

Fig IV.3: APA-2 stepsize parameter μ (λ=0.95)

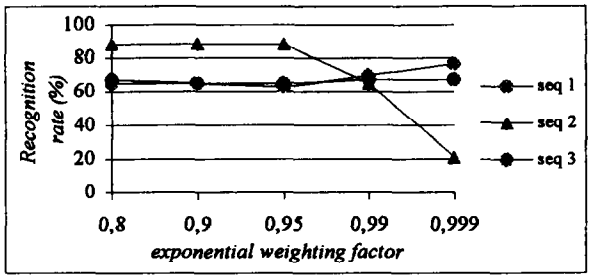

Fig IV.4 : APA-2 exponential weighting factor λ (μ =0.1)

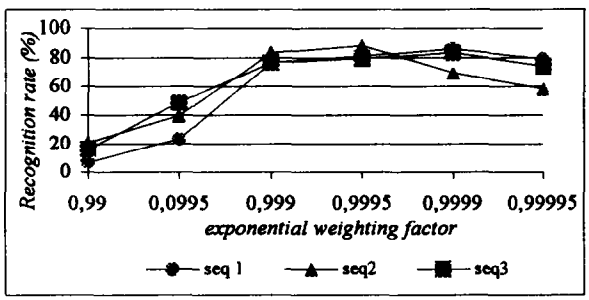

Fig IV.5 : FQR_LS exponential weighting factor λ

the MDFO and FQR-LS algorithms are quite robust to the choice of their parameters, figures (IV.2-IV.5).

IV.3.3 Recognition rate comparison in noise conditions

Figure (IV.6) shows the recognition rates of each stereophonic echo cancellation algorithm on a base of 18 radio sequences for 2 different speakers, one female and one male. 1548 words had to be recognised. The parameters of each algorithm were chosen to give the best recognition rates on the above mentionned sub-database : NLMS μ=0.15 ; APA-2 μ=0.1, λ=0.95 ; MDFO μ=0.3 ; FQR-LS λ =0.9995. NLMS and APA-2 performances slightly decrease in noise conditions. This can be explained by their high variance in recognition rates according to radio sequences (var : NLMS 7,74% ; APA-2 8,53%). MDFO and FQR-LS algorithms seem to be more robust to the high diversity of radio sequences used, almost reaching 90% of recognition rates (var : MDFO 5.7% ; FQR-LS 4.2%).

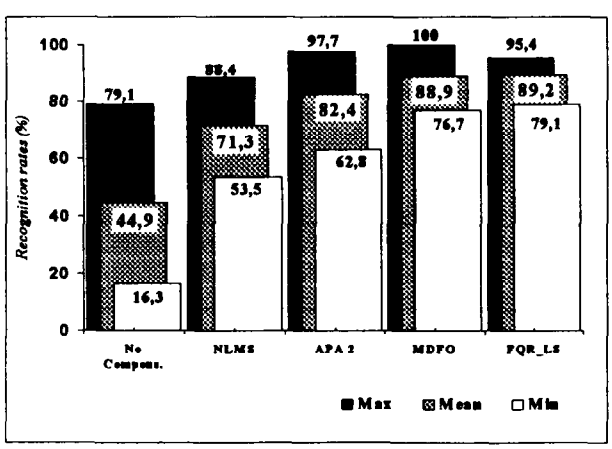

Fig IV.6 : Recognition Rate Comparison in noise conditions.

IV.3.4 Subband structure with FQR-LS Algorithm

As suggested in [5], we propose to reduce the overall complexity of the stereophonic echo canceller based on the FQR-LS algorithm by using a subband structure. The results obtained on the previously defined sub-database (2 speakers, 18 radio sequences, SNR1 = 0 dB, SNR2 = -5 dB) are given in figure (IV.7). We have used a perfect reconstruction modulated filter bank, based on Extended Lapped Transform (ELT), [13]. The length of the basis function is L=2KM, where M is the number of subbands (M = 4, 8, 16, 32, 64), and K is the overlapping factor, equal to 4. The length of the subband adaptive filters matches the 256 taps full-band adaptive filter : 4 taps adaptive filters have been used in the 64-subbands structure.

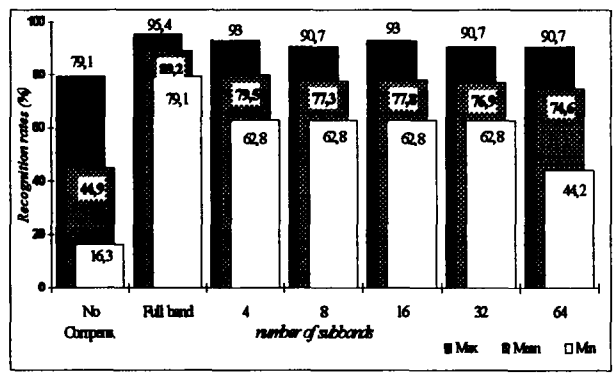

Fig IV.7 : Subband structure with FQR-LS Algorithms.

No specific optimisation has been performed according to each subband. We have not incorporated any cross-band adaptive filters, and we did not use any oversampling scheme. Even though the results are degraded when using the 4-subbands structure when compared to the full-band FQR-LS , we can note that the performances are maintained up to the 32-subbands structure, almost performing as the full-band APA-2 algorithm. We expect to greatly improve these results by incorporating independent controlled convergence of the subband adaptive filters, [6]. We also think that part of the degradations is due to aliasing, which seems to be more critical for some music sequences than for speech signals. This can be tackled with cross-band filters or oversampling schemes.

IV.4. Comparison of recognition sytems performances

IV.4.1 Test condition.

The two recogniser systems were run on the same sub-database used in IV.3.3 (2 speakers, 18 radio sequences, SNR1 = 0 dB, SNR2 = -5 dB). Stereophonic echo cancellation was performed using MDFO algorithm. As mentionned in IV.1.1, the words are pronounced in quiet conditions and are well separated by silence pauses. So, it is interesting to compare the performances of the two recognisers in an isolated word application. We give as reference the recognition rates computed using the theoretical labels (cf IV.2), ie those obtained when not taking into account the problems of Voice Activity Detection.

IV.4.2 Evaluation of the system.

The figure (IV.8) summarises the performances of these systems when used with or without car-radio noise compensation . We define the following notations : IWT for the isolated word recogniser using the theoretical labels, IWO

for the isolated word recogniser using a VAD, WS for the word-spotting recogniser.

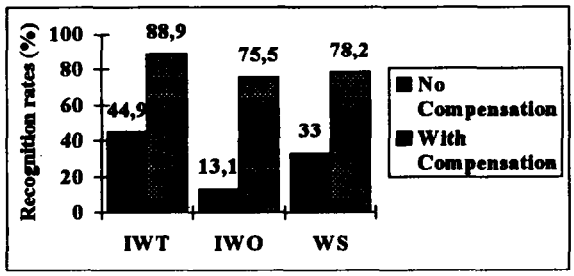

Fig IV.8: Recognition without/with car-radio noise compensation

We can note that the performances obtained with the word-spotting recogniser are better than those with the isolated word recogniser, and this whether car-radio noise compensation is performed or not. The concern of using a word-spotting recogniser appears here for such applications: it enables the system to recognise the word sequences which have been detected as single word by the VAD.
However, we still don't achieve the performances obtained by the isolated-word recogniser when word insertions and elisions due to the Voice Activity Detector are corrected (use of theoretical labels).

V. CONCLUSIONS

We have shown the efficiency of stereophonic acoustic echo cancellation for car-radio noise compensation in the context of speech recognition for voice-dialing or voice-controlled functions in a car. These results have been obtained on a database representative of the different encountered conditions, showing the performances of classical algorithms (NLMS, APA-2, MDFO, FQR-LS). Some modified algorithms for stereophonic echo cancellation are now under study, either for transmission or speech recognition applications.
As for recognition system, no specific optimisation was performed compared to a standard speech recognition system used in noisy car environment, which also leaves room for performances improvement.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Y.Grenier and E.N.S.T., for providing technical supports for recordings of a preliminary database, C. Louis and O. Rigaut for their helpful works.
Part of these works are supported by the A.N.R.T. under grants N° 167/95 and N° 115/97.

V. REFERENCES

[1] Alexandre P., Lockwood P., 1993, *"Root Cepstral Analysis: A Unified View. Application To Speech Processing In Car Noise Environments"*, Speech Communication 12, pp 277-288, 1993.

[2] Amand F., Benesty J., Gilloire A., Grenier Y., 1995, *"Un algorithme d'annulation d'écho stéréo de type LMS prenant en compte l'inter-corrélation des entrées"*, 15 ème Colloque GRETSI - Juan-Les-Pins, 18-21 septembre 1995, pp 407-410, Septembre 1995.

[3] Bellanger G., Regalia P., 1991, *"The FLS_QR algorithm for adaptative filtering : The case of multichannel signals"*, Signal Processing 22, pp 115-126, 1991.

[4] Boudy J., Capman F., Lockwood P., 1995, *"A Globally Optimised Frequency-Domain Acoustic Echo Canceller For Adverse Environment Applications"*, 4th International Workshop on Acoustic Echo and Noise Control, Roros, Norway, pp 95-98, June 1995.

[5] Capman F, Boudy J., Lockwood P., 1995, *"Acoustic Echo Cancellation Using A Fast-QR-RLS Algorithm and Multirate Schemes"*, IEEE ICASSP-95, Detroit, pp 969-972, 1995.

[6] Capman F, Boudy J., Lockwood P., 1997, *"Controlled Convergence Of QR Least-Squares Adaptive Algorithms - Application To Speech Echo Cancellation"*, IEEE ICASSP-97, pp 2297-2300, Munich, 1997.

[7] Dufour S., Glorion C., Lockwood P., 1996, *"Evaluation of the Root-Normalised Front-End (RN_LFCC) for Speech Recognition in Wireless GSM Network Environments"*, IEEE ICASSP-96, pp 77-80, 1996.

[8] Glanz M., Linhard K., Kroeschel K., 1990, *"Speech Recognition in Cars with Noise Suppression and Car Radio Compensation"*, 22nd ISATA, Florence, 14-18 mai 1990, pp 509-516, May 1990.

[9] Lockwood P., Boudy J., 1992, *"Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars"*, Speech Communication, Vol. 11, Nos 2-3, pp 215-228, June 1992.

[10] Makino S., Strauss K., Shimauchi S., Haneda Y., Nakawaga A., 1997, *"Subband Stereo Echo Canceller Using The Projection Algorithm with Fast Convergence to the True Echo Path"*, ICASSP-97, pp -, 1997.

[11] Sondhi M.M., Morgan D.R., Hall J.L., 1995, *"Stereophonic Acoustic Echo Cancellation - An Overview of the Fundamental Problem"*, IEEE Signal Processing Letters, Vol. 2, n° 8, pp 148-151, August 1995.

[12] Malvar H.S., *"Signal Processing with Lapped Transforms"*, Artech House.