SPEECH RECOGNITION SYSTEM FACING A CAR RADIO

Fabrice Amand (1), Ivan Bourmeyster (2), Giorgio Parladori (3)

(1) Cefriel, Via Emanueli 15, 20126 Milano, Italy
(2) Alcatel Mobile Phone, 32 av. Kleber, 92707 Colombes Cedex, France
(3) Alcatel Advanced Technologies, Via Trento 30, 20059 Vimercate (MI), Italy

ABSTRACT

This paper deals with a speech recognition system in a car facing a car radio. The speech recogniser can be used either for a mobile phone or for a board computer. So, it will be possible to activate the mobile phone together with the car radio. It will also be possible to voice control the board computer in the same context (car radio active). The sound diffused by the car radio loud-speakers disturbs the speech recogniser [1]. To cancel the loud-speakers echo picked up by the microphone, we show and evaluate the use of an Acoustic Echo Cancellation technique in the Stereophonic case (AECS).

1. INTRODUCTION

Our final objective is to get the same speech recognition system behaviour with or without the car radio on. This objective can be reached if we cancel the car radio echo. To cancel the car radio echo, we propose to use the well known Acoustic Echo Cancellation technique (AEC). In this car radio context, **the AEC reference input signals are not only speech** (for example an information radio) **but also music** (as jazz, rock..). We know from [2] that all the AEC algorithms behaviours are mostly identical either with music or with speech. We confirm it. Furthermore, as the number of loud-speakers can be equal or greater than two, we consider an **Acoustic Echo Cancellation** problem but in the **Stereophonic** case **(AECS)** [3]. In this context, we apply the AECS system described in figure 1 to the speech recogniser. In the following, we give a first evaluation of the usefulness of the AECS system, without theoretical result.

2. DATA BASE DESCRIPTION

The reference input signals have been recorded from CD and FM radio. They consist of jazz, rock and classical music and speech. Then, they have been diffused in a car with the two loud-speakers fitted out in the front of the car. During the diffusion, the

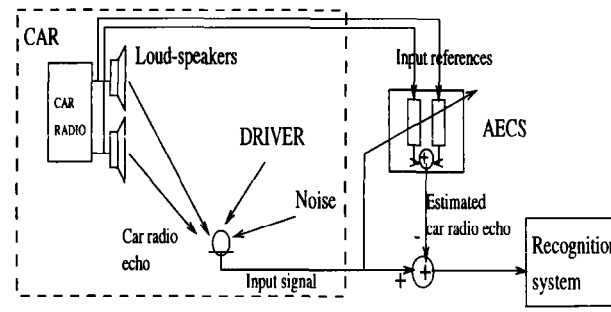

Figure 1: AECS system.

driver had to speak to simulate a vocal command. The coherence function (*gama*) between the left and the right loud-speaker inputs ($x1$ and $x2$) is:

$$gama(f) = \frac{S_{x1x2}(f)}{\sqrt{S_{x1x2}(f) S_{x1x2}(f)}} \quad (1)$$

where:

$$S_{x1x2}(f) = \sum_{r=-\infty}^{r=\infty} E[x1(n)x2(n-r)]e^{-i2\pi fr} \quad (2)$$

In the figure 2, we show two coherence functions of two input reference files issued from an information radio (a french radio diffusing news), called *information*, and from a classical music radio, called *classical*. We can see, right part of the figure 2, that the coherence function of the *information* left and right input signal references is nearly equal to 1 (that means that the left and right input references are almost identical). That is not the case for the *classical* signal, left part of the figure 2, where the coherence function values are very different. Here, we point out a critical situation.

3. CRITICAL SITUATION

We know from [3] that for identical AECS algorithms input references, the algorithms (using two adaptive filters to identify the two real impulse responses) could be not able to identify the real impulse responses.

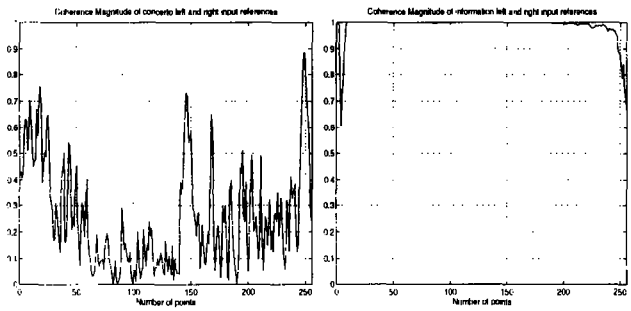

Figure 2: Coherence function of the two input reference files: *Classical* (left) and *Information* (right).

This is the case for *information* where the coherency function is nearly equal to 1, see figure 2. With *information* as input and after reaching the asymptotic convergence, we have verified that the each adaptive filter identifies a mix of the real impulse responses. In such a situation, we have simulated a scanning from the information radio to the concerto radio. With *concerto*, due to the difference of the left and the right loud-speakers, see figure 2, the two adaptative filters converge to the real impulse responses. So, at one moment the acoustic path changes heavily introducing an important residual error. A very sensitive algorithm as an APA algorithm can diverge. If such a case occurs, we obviously understand that the speech recogniser will not work.

4. THE SPEECH RECOGNITION SYSTEM

The block diagram figure 3 shows the main processing modules of the whole speech recogniser.

Front end processing: With reference to figure 3, the following step are performed:

- Digitised speech, sampled at 8kHz, is preemphasized and windowed.
- After FFT, a MEL extration is performed. 30 coefficients over MEL scale are calculated from the square modulus of the 256 FFT coefficients.
- *Noise Estimation (NE):* noise estimation is updating of each MEL band and is in the form:

$$N(t) = alpha * N(t-1) \quad (3)$$

where *alpha* is function of signal to noise ratio. The *NE* initialization period is 44 trames. In this equation, we can see that the noise model is a white noise model. So, we can expect that the car radio (diffusing music or speech) will be not taken into account by the *NE*.

- *Maximun Likelihood Coefficients (MLC)* estimation: for each band a parameter is obtained as function of the signal to noise ratio of that band, according to the maximum likelihood function with soft decision scheme algorithm:

$$MLC(i) = 0.5 * (1 + \sqrt{\frac{S(i)-N(i)}{S(i)}}) * P[H_1/S(i)] \quad (4)$$

Where $P[H_1/S_f]$ is the probability of speech, given the observed power $S(i)$ in the i-th band. $P[H_1/S(i)]$ is a function of the signal to noise ratio too. These parameters can be considered as a measure of the speech probability in each MEL band. They allow an integration of operation of VAD in the algorithm.

- *Voice Activity Detector (VAD):* The MLC of each band is filtered in the time domain. It is used to determine the state (word / no word) of the associated mel-band by comparing it with a threshold.
- *DCT:* 12 coefficients from the 30 MEL coefficients as calculated and used by DTW for the matching between references and utterances.
- *Liftering:* a sine square liftering and a normalization are applied to the 12 DCT coefficients.

In our case, we do not consider the MFCC coefficient power. So, for an input signal we can give the associated VAD and the 12 MFCC coefficients calculated for each 120 samples at 8 kHz.

Dynamic Time Warping or Hidden Markov Model: Depending of the application, we can use DTW or HMM algorithm [1]. The algorithm used is started and stopped by the VAD signal.

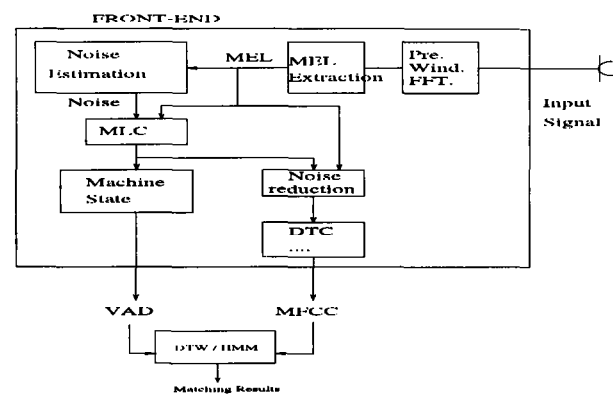

Figure 3: Speech Recognition system.

5. AECS EVALUATION WITH THE SPEECH RECOGNISER

5.1. Simulation context description

In our simulations, we used only the Front End part that delivers the Vocal Activity Detection (VAD) and the Mel Frequency Cepstrum Coefficients (MFCC). We make the hypothesis that if the VAD and MFCC coefficients are the same with or without the car radio

on, we obtain the same speech recognition behaviour. In the following simulations, we show the VAD obtained by the speech recogniser with the signal picked up before and after the AECS processing (we add a star to show the vocal command VAD obtained with the driver vocal command only). We also show the cepstral distance between the MFCC coefficients before and after the AECS processing with reference to the driver vocal command, see figure 4. The driver vocal command consists of three french words: 'vitre avant' 'vitre arriere' and 'fermer'. We show results obtained with three radio stations: *information* issued from a french radio news, *Rock* and *Concerto*.

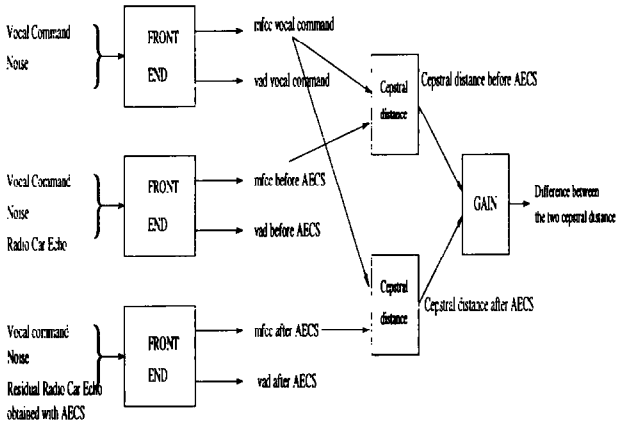

Figure 4: Methodology of evaluation.

The cepstral distance is calculated only when the driver speaks:

$$CepstralDistance(i) = \sum_{j=1}^{j=12} (MFCC_{1,j}(i) - MFCC_{2,j}(i))^2 \quad (5)$$

The VAD used for the AECS is obtained by listening to the input signals. We use the two-channel LMS algorithm [3] with the adaptation step fixed to $mu = 0.2$ and the size of the adaptive filter fixed to $L = 256$. The average Echo Return Loss Enhancement (ERLE) between car radio echo before and after AECS processing is about 20 dB. To avoid problems due to the Front-end initialization the simulation data consists of the input sequences repeated twice.

5.2. Simulation results

The VAD (figures 8, 9, 10) obtained before the AECS processing is totaly different from the VAD obtained on the vocal command. We note that the VAD before AECS has the same characteristics as the VAD obtained on the car radio echo. **We understand now why in [1] the speech recogniser can not work with the car radio!** The VAD obtained after the AECS processing is nearer to the driver vocal command VAD. But, we remark that some VAD cases of false detection, non detection and too long detection remain. The AECS VAD obtained by listening is different from the Front-End VAD. For the Front-End VAD 'vitre avant' is a unique word to be recognized, it is two words for the AECS VAD. If we consider that the filter adaptation introduces a perturbation for the speech recogniser, the filter adaptation has to be stoped during these transitions. So, the Front-End VAD can also be used by the AECS system. Though, we know the risks of a looped system.

For the MFCC coefficients (figures 5, 6, 7) the cepstral distance between the signal picked up after AECS and the vocal command signal is smaller than the cepstral distance between the signal picked up before AECS and the vocal command signal. If the MFCC coefficients after AECS are nearer to the MFCC driver vocal command coefficients, we can expect a better recognition rate. **So, we can conclude on these simulations that the AECS will improve the recognition system behaviour.** We point out that the gain (difference between the two cepstral distances) can have negative values.

6. CONCLUSION

We have given a first evaluation of an AECS system applicability with a speech recognition system in the presence of a car radio. The AECS system improves the VAD with respect to the absence of AECS processing. Furthermore, the cepstral distance between the signal picked up after AECS and the vocal command signal is smaller than the cepstral distance between the signal picked up before AECS and the vocal command signal. But we conclude that the AECS processing used, which gives 20dB of echo attenuation, seems to be not sufficient to keep the same recognition rate with or without the car radio.

References

[1] Ivan Bourmeyster and Al., 'Voice controlled mobile phone for car environment', Eusipco96, Trieste (Italy), September 1996.

[2] Frank Scheppach and Al., 'Echo compensation and noise suppression for speech recognition applications', Eusipco96, Trieste (Italy), September 1996.

[3] Fabrice Amand, 'Etude de l'annulation d'echo multivoie et application a la teleconference de haute qualite', PhD, FT-CNET Lannion, France, March 1996.

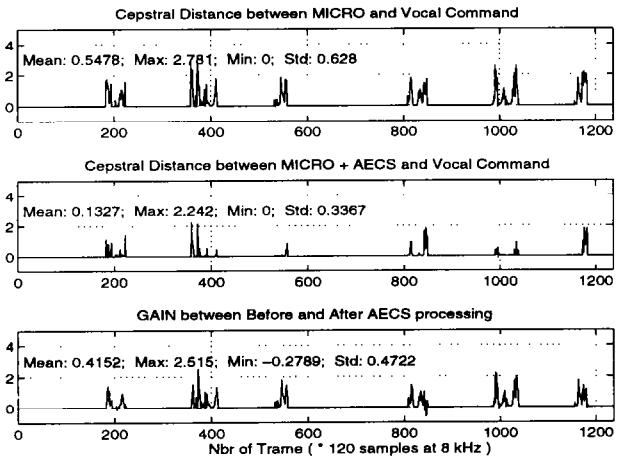

Figure 5: Cepstral MFCC distance obtained with *Information*.

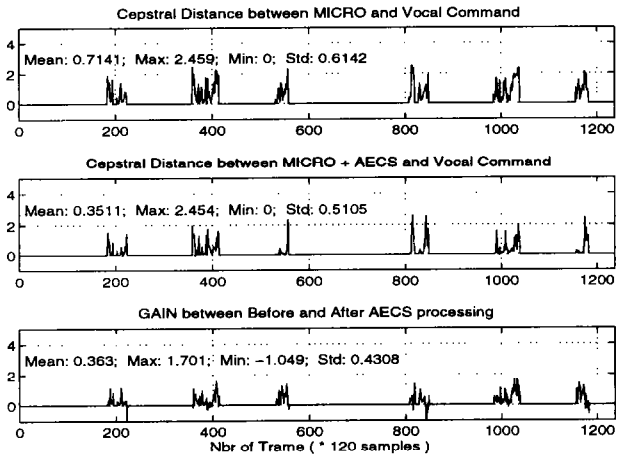

Figure 6: Cepstral MFCC distance obtained with *Concerto*.

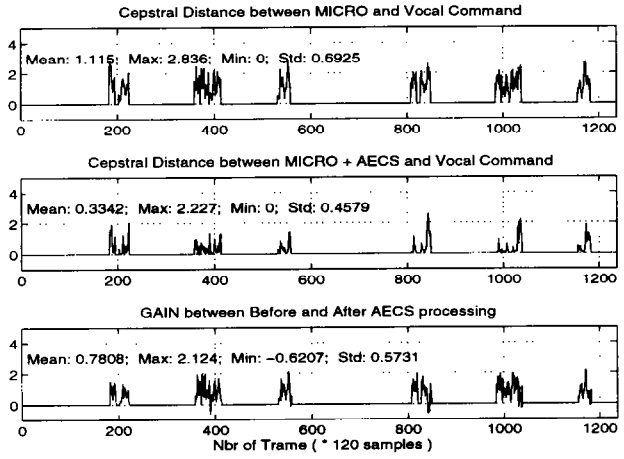

Figure 7: Cepstral MFCC distance obtained with *Rock*.

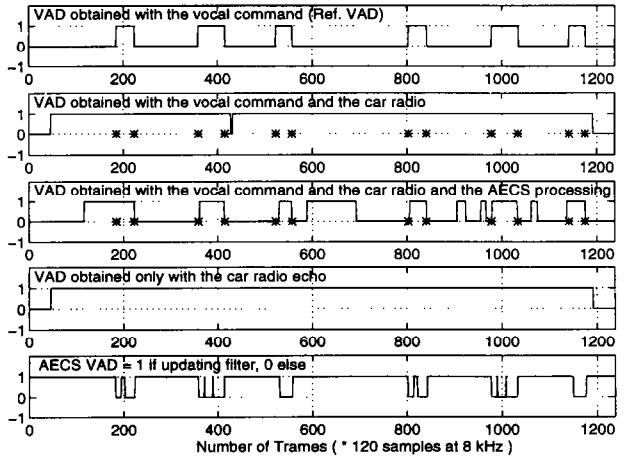

Figure 8: VAD obtained with *Information*.

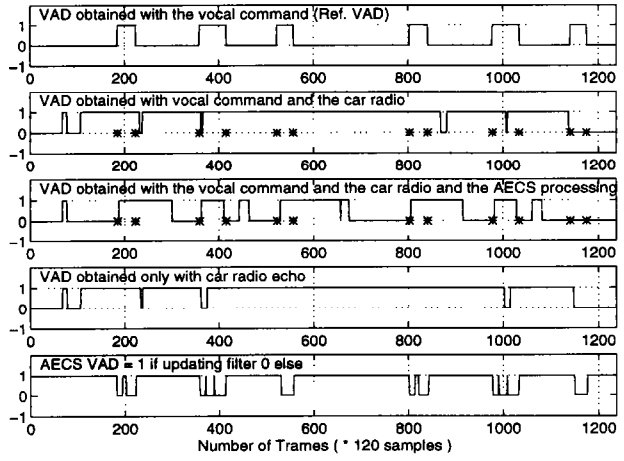

Figure 9: VAD obtained with *Concerto*.

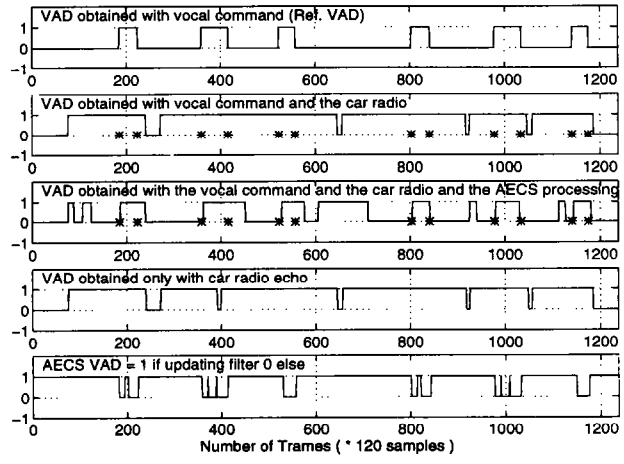

Figure 10: VAD obtained with *Rock*.