A new Algorithm for double talk detection and separation in the context of digital mobile radio telephone

Hassan EZZAIDI *, Ivan BOURMEYSTER**, Jean ROUAT *

* D.S.A, Université du Québec à Chicoutimi, Québec , CANADA, G7H 2B1
**Alcatel Mobile Phones 32, avenue Kléber 92707 Colombes Cedex FRANCE
hezzaidi@uqac.uquebec.ca , ivan.bourmeyster@art.alcatel.fr , jrouat@uqac.uquebec.ca

I - ABSTRACT

The paper describes a new technique that enhances the Voice Activity Detection (V.A.D) performance between the remote speaker (received signal) and the local speaker (located in the vehicle) in the context of mobile radio telephone environment. We use an Auditory Pitch and voiced/unvoiced Detection (A.P.D) algorithm in conjunction with an Auto Regressive (A.R) analysis in order to remove the remote speaker's voice signal from the car hands-free microphone signal. Results are compared with the reference system that doesn't include the APD.
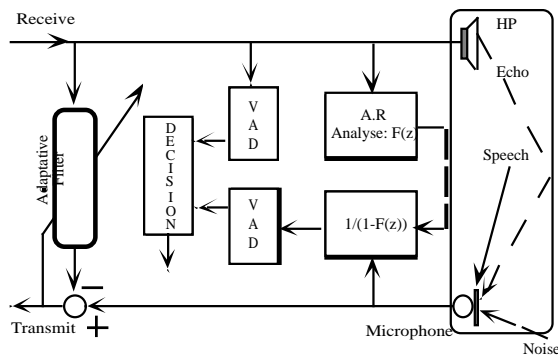
II - INTRODUCTION

figure 1: Reference Structure

The radio-telephone hands-free environment is characterised by a corrupted and additive noise [Degan, 1988], as well as by an echo phenomenon that is due to the coupling between the loudspeaker and the microphone. Consequently, Noise Reduction (N.R) and Acoustic Echo Cancellation (A.E.C) must be included in the phone equipment.

Until now contributions considered this particular problem by proposing various algorithms for N.R [Windrow, 1975] [Lim, 1979] [Boll, 1979] [Ephraim, 1984], and A.E.C [Macchi, 1988] [Ozeki, 1984] [Haykin, 1991].
Generally, The A.E.C is an adaptive filter which coefficients must be updated only when the remote speaker is talking and the local one does not. The technique used is also referred to as double-talk detection. Usually, this discrimination is performed by algorithms based on power signal considerations.

We study two techniques exploiting the pitch information that seem to be reliable regarding non linear distortion introduced in the car.

III - ANALYSIS AND METHODOLOGY

We have simulated the V.A.D module by files, where the discrimination between the two speakers has been evaluated manually, in order to have a robust reference system. They will be referred to as VadRx(n) and Tx(n) (= 1 or 0 depending on vocal activity).
We did not use any N.R module. The speech input sampling frequency is fixed to 8 KHz.

A) Database

The database was recorded in real conditions in a car for various situations :

- Speed : 0, 60, 90, 130 km/h;
- Windows : open or closed;
- Speakers : men or women;

The database is a series of couple of files : (received signal, transmitted signal) recorded synchronously.

B) Two channels pitch detection paradigm

The technique consists in replacing directly the VAD in figure1 by the APD. The APD is then used to decide on the voice activity of both speakers on the receive and microphone channel. We analysed the remote speaker's signal in conjunction with the microphone signal. Unfortunately, when the pitch estimate of the two speakers are close or when the noise becomes too high on the local side, we could not find any good criteria to distinguish or correlate them.
We have tried a pseudoresidual technique [Prado, 1993] which comprises an inverse filtering of the microphone signal by using the auto regressive analysis performed on the remote speaker 's signal. The analysis was carried out with that so called pseudoresidual signal in place of the microphone signal. In that case, better results were achieved in all noisy conditions for our database: the APD was not detecting any voice activity during only echo phases. Good pitch detection was achieved when the local speaker is active. However, in clean conditions (car stopped), pitch was detected for both echo and local speaker.

Because of the computational complexity generated, no further investigation was made on double-detection technique.

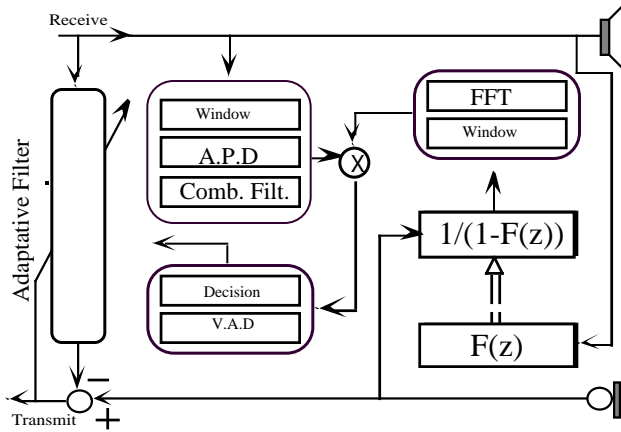## C) Single channel pitch detection paradigm

figure 2: Structure(1,2,3)

The technique refers to the use of the pitch estimate over the remote speaker signal (see result of estimated pitch in figure 3). First the remote speaker vocal track contribution is estimated through a classical LPC method (formant predictor). From the microphone signal we compute the pseudoresidual signal. A great reduction of the remaining echo of the vocal track from the remote speaker is achieved. Then, the pitch estimation is used to synthesise a comb filter in the frequency domain. Concurrently, we calculate the Fast Fourier Transform on the pseudoresidual signal. For each frame a multiplication with the frequency transfer function of the comb is performed. The accumulation leads to an energy estimate that is finally processed by a VAD detection module. That last step gives a decision on the local speaker voice activity (see figure 2).

According to [Kabal, 1989], the cascade Formant-Pitch predictor outperforms the Pitch-Formant predictor. We have considered in this analysis 5 structures as follow:

structure 1 :
    The formant cancellation is performed before the pitch cancellation .

structure 2 :
    The pitch cancellation is performed before the formant cancellation.
 structure 3 :
    First, a pitch is performed on a microphone signal, then the formant cancellation follow as structure 2 and finally another pitch cancellation is performed.

We also considered structures 1 and 2 with the difference that the pitch estimate and the LPC analysis were carried on using the output of a standard NLMS adaptive filter rather than on the remote speaker signal. The adaptive filter was mounted as for a classical acoustic echo control system. We labelled these structures 4 and 5.

Moreover, different techniques were investigated to suppress the frequency components from the pitch harmonics locations :

 H1: remove the frequency components closest to a pitch harmonics.
 H2: remove the frequency components framing a pitch harmonics.
 H3: a weighting of the frequency components closest to the pitch harmonics is performed.

The weighting factors are calculated using the distance between the frequency component and its closest pitch harmonics.
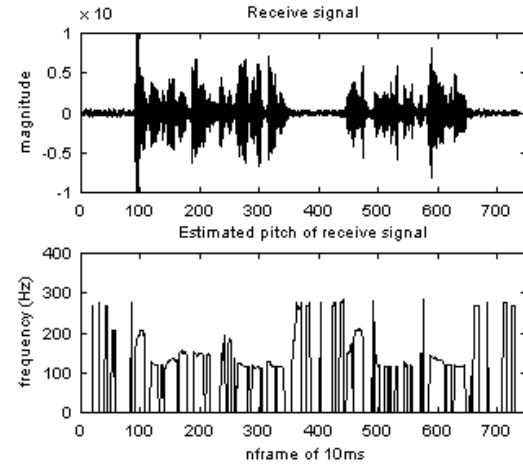

figure 3: x-axis values are given in (Cs). a) signal of remote speaker. b) value of estimated pitch at rhythm of 10 ms.

D) Criteria for evaluation

We used the Alcatel R.L.E (in French : Rapport Local à Echo) criteria defined as the log ratio of the local speaker energy to the echo energy measured on the microphone signal during the remote speaker vocal activity phases.

During these phases, two situations have to be discriminated at the microphone signal $Tx(n)$ level : double-talk situation $(DbTlk(n)=1)$ and echo alone $(DbTlk(n)=0)$.

The R.L.E measures the discrimination, and is calculated as follows :

$$RLE = 10 * \log_{10}\left(\frac{EnLoc}{EnEcho}\right)$$

$Where:$

$$EnLoc = \frac{\sum_{n}^{N}(Tx(n))^2 * VadRx(n) * (1 - DbTlk(n))}{\sum_{n}^{N} VadRx(n) * (1 - DbTlk(n))}$$

$$EnEcho = \frac{\sum_{n}^{N}((Tx(n))^2 * VadRx(n) * DbTlk(n))}{\sum_{n}^{N}(VadRx(n) * DbTlk(n))}$$

E) RESULTS

The RLE mean measure is reported in tables 1 to 3 for various driving conditions. At 0 km/h the engine of the car is running. The 60 km/h recording are mostly characterised by city roads. The car windows were closed for all results presented in table (1,2,3).

H1, H2 and H3 characterise the harmonic suppression, a description is given in section (C) for harmonic methods and the used structures.

The RLE gain is calculated by the difference between the proposed structure RLE and the reference structure RLE (see figure 4 and figure 5).

Table 1: results of structure 1

	H1	H2	H3
0 km/h	1.31	1.90	2.05
60 km/h	0.46	0.96	0.82
90 km/h	0.23	0.44	0.95
130 km/h	1.091	0.568	0.24

Table 2: results of structure 2

	H1	H2	H3
0 km/h	1.21	1.56	2.02
60 km/h	0.22	0.53	0.77
90 km/h	0.75	0.31	0.95
130 km/h	0.51	0.22	0.24

The structure 1 outperforms the structure 2 at 0 km/h and 130 km/h where the background noise and coupling between the loudspeaker and microphone seems slightly changed.
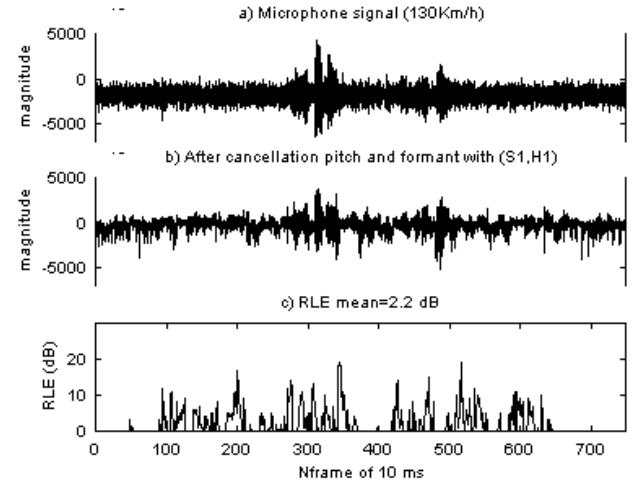In the contrary, the structure 2 outperforms the structure 1 at 0 km/h and 90 km/h (see figure 4).

figure 4. x-axis values are given in (Cs). a): microphone signal (130 km/h, male speaker). b): microphone signal after cancellation of formant pitch using structure 1 and cancellation harmonic method H1. c): RLE performance shows the enhancement discrimination for the local speaker with this proposed implementation .

Table 3 : results of structure 3

	H1	H2	H3
0 km/h	2.90	3.53	3.5
60 km/h	3.33	4.13	4.5
90 km/h	4.00	5.31	5.0
130 km/h	4.50	4.7	4.0

The structure 3 gives the bests results with a RLE criteria 3 dB better for all conditions driving.

The harmonic suppression H3 seems to be the best cancellation pitch method. Probably, because it compensates for the limited frequency resolution.

Structure 4 and 5, obviously degrade the pitch estimation, probably because the coefficient of the adaptive filter keep changing with time. Moreover, this structures create a closed loop in the system since the adaptation decision is feedback into the adaptive filter module. A more extensive study showed to made in order to guaranty that adaptive filter divergence never occurs.
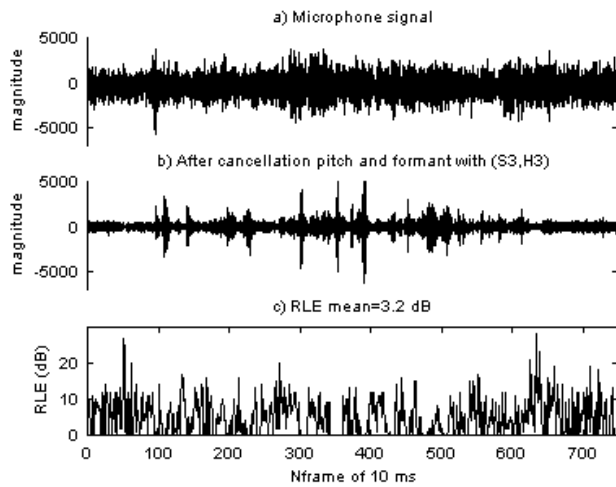
figure 5. x-axis values are given in (Cs). a): microphone signal. b): microphone signal after cancellation of formant pitch using structure 3 and cancellation harmonic method H3. c): RLE performance (dB).

IV- CONCLUSION

Various techniques were investigated to introduce pitch estimation for solving the double-talk detection problem in the car hands-free context. A double pitch detection technique on both received and transmitted signal was studied. Unfortunately pitch coherence or mismatch was difficult to achieve in many noisy conditions.

A single pitch detection principle was investigated. It consists in enhancing the local speaker to echo discrimination on the microphone signal. The enhancement is achieved through a vocoder-like analysis of the remote speaker signal used to compute a residual-like (long term and short term) on the microphone signal.

An evaluation criteria was proposed to assess the discrimination performances of various structures and system configurations : different locations in the system where to apply the effects of A.R predictor and pitch estimate were studied. The introduction of pitch estimation always improves the discrimination criteria. Experiments were run on a database recorded in various car environment situations and various speakers.

The introduction of noise reduction was not studied but seems to be an interesting area for further investigation. The various stages for computing the residual signal could be combined with frequency domain noise reduction
technique . The noise reduction could be tuned so that residual echo is considered as noise. It would then be attenuated leading to even better discrimination capabilities.

REFERENCE

N. D. Degan and C. Prati (1988), "*Acoustic noise analysis and speech enhancement techniques for mobile radio applications*", Signal Processing 15, pp 43-56.

R. P. Ramachandran, Peter Kabal(1989), "Pitch Prediction Filters in Speech Coding", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP 37, no. 4. April.

B. Widrow et al. (1975), "*Adaptive noise cancelling: principles and applications*", Proc. of the IEEE, Vol. 63, No. 12, pp. 1692-1716.

J.S. Lim, A.V. Oppenheim (1979), "*Enhancement and bandwidth compression of noisy speech*", Proc. of the IEEE, Vol. 37, No. 12, pp. 1586-1604, December.

S.F. Boll (1979), "*Suppression of acoustic noise in speech using spectral subtraction*", IEEE Trans. on ASSP, Vol. 27, No. 2, pp. 113-120, April.

Y. Ephraim, D. Malah (1984), "*Speech enhancement using a minimum mean square error short time amplitude estimator*", IEEE Trans. on ASSP, Vol. 32, No. 6, pp. 1109-1121, December.

O. Macchi, M. Bellanger (1988), "*Le filtrage adaptatif transverse*", Traitement du Signal, Vol. 5, No. 3, pp. 115-132.

K. Ozeki, T. Umeda (1984), "*An adaptative algorithm using an orthogonal projection to an affine subspace and its properties*", Electronics and Communications in Japan, Vol. 67-A, No. 5, pp. 19-27.

S. Haykin (1991), "*Adaptive filter theory*", Second Edition, Prentice-Hall, Englewood Cliffs, New Jersey.

J. Prado, E. Moulines (1993), "*Frequency-domain adaptive filtering with application to echo cancellation*", Proc. of the Third International Workshop on Acoustic Echo Control, pp. 249-258.