COMPARISON BETWEEN SUBBAND AND FULLBAND NLMS FOR IN-CAR AUDIO COMPENSATION AND HANDS-FREE SPEECH RECOGNITION

Maurizio Omologo, Christian Zieger

omologo/zieger@itc.it
ITC-irst, Povo, via Sommarive 18, 38050, Trento, Italy

ABSTRACT

This paper deals with audio compensation in the car environment for the development of a hands-free spoken dialogue system with barge-in functionalities. While in a preliminary work we investigated the problem, given an interfering radio signal to compensate, here we focus only on the compensation of speech audio generated by a text to speech synthesizer. The latter one is a more difficult signal to manage, as speech is colored and non stationary or quasi stationary and this degrades the performance of the AEC if a simple NLMS is used. In this paper we will introduce a Subband Acoustic Echo Cancellation (SAEC) for compensating the synthetic prompt speech and we will compare it with a Fullband Acoustic Echo Cancellation (FAEC) demonstrating the good effectiveness in terms of speed of convergence, robustness against noise and computational complexity. The system performance are being measured in terms of Word Error Rate % (WER) by recognizing isolated and connected digits.

1. INTRODUCTION

In a preliminary work [1] a fullband NLMS based compensation system was investigated to deal with speech recognition given an interfering in-car radio signal. NLMS was applied on signals belonging to real Italian SpeechDatCar database [1], where the signals were acquired under various environmental conditions. The time available to the system for convergence, here denoted as t_s and corresponding to the time interval between the radio signal start time and the user speech start time, was large enough to allow a good convergence of NLMS.

The purpose of that work was a preliminary study for the introduction of a barge-in functionality in a hands-free spoken dialogue system. The barge-in functionality is the possibility the user has to interrupt the system during the prompting of a voice message. For this purpose it is necessary to remove the acoustic echoes of the synthetic prompt speech from the far microphone signal. For this reason, in this work we focus only on the audio compensation of the synthetic speech, that is we investigate the impact of different levels of background noise as well as the impact of t_s on a fullband and subband NLMS. The system performance is measured in terms of WER %. The two AEC systems are applied on a simulated database, which is characterized by different values of SNR and SIR.

In the next section we describe SAEC and the method to control the step size for each subband. In section 3 a comparison between SAEC and FAEC in terms of speed of convergence and computational complexity is made. Section 4 refers to the HMM recognizer used for the experiments. Section 5 reports on the development of an artificial database used for the experiments and shows the relative results for FAEC and SAEC. Finally the conclusions and the possible future work are reported in section 6.

2. SUBBAND ACOUSTIC ECHO CANCELLATION

In a subband acoustic echo canceller the reference and the far microphone signals are split in M subbands through the Analysis Filter Bank (AFB), then NLMS is applied to each subband independently and at last the signal is reconstructed through the Synthesis Filter Bank (SFB) [2]. In general the down-sampling factor K in the AFB is less/equal to the number of subbands M. If $K = M$ the filter bank is called critically sampled, otherwise ($K < M$) it is non-critically sampled. The filter bank is derived by the frequency “sliding” of a lowpass prototype filter of length L_p ($L_p = 320$ in our experiments).

The downsampling in the AFB causes aliasing which is suppressed in the SFB with a proper choice of the prototype filter. Aliasing is a disturbing signal for the NLMS convergence process that can degrade seriously the system performance [3]. Critically sampled systems make use of cross-terms between adjacent bands to cancel the aliasing [4], while for a non-critically sampled system, the aliasing can be reduced by decreasing the downsampling factor. In fact by decreasing K the distance between the bands increases, while the subband width remains the same.

As computational load can be saved by using an efficient implementation of the filter bank, here we adopted that reported in [5], where the polyphase decomposition is exploited to remove redundancy computations and the FFT algorithm is used for the modulation. If the reference signal is real, in order to save more computational load, it is possible to apply NLMS only on the first $M/2 + 1$ subbands and then exploit the hermitian simmetry in the frequency domain to reconstruct the full band signal.

To estimate the system impulse response in each subband it is possible to apply the NLMS algorithm with a control of the step-size parameter based on the method called “delay coefficients” [6, 7, 8], which was used in other experiments for FAEC [1]. The subband step size control is described in the following section. The approach of controlling the step-size parameter in each subband was also adopted in [9], although in that case a different technique was used to compute the optimal step-size for each subband.

2.1. Subband step size control

As the outputs of the AFB are in the complex domain the NLMS learning rule for complex signals must be used [2]

$$\mathbf{h}_i(k+1) = \mathbf{h}_i(k) + \mu_i(k)\frac{e_i^*(k)\mathbf{x}_i(k)}{\|\mathbf{x}_i(k)\|^2} \qquad (1)$$

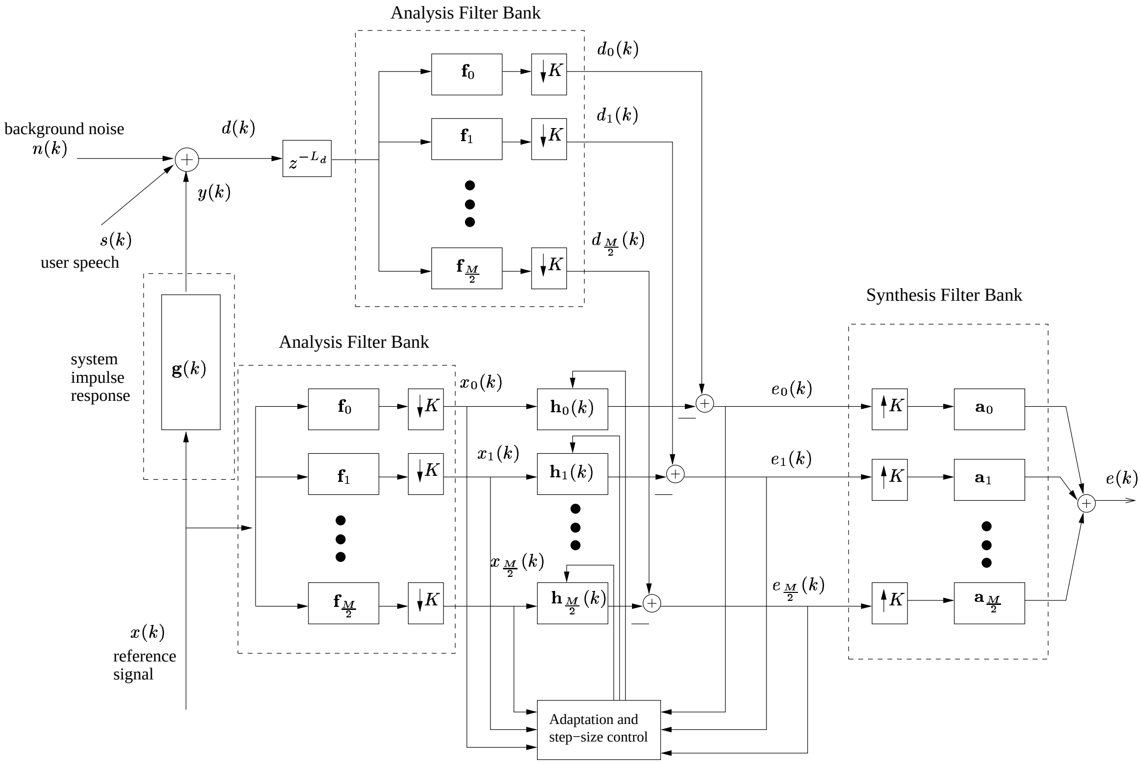


Figure 1: System block diagram for SAEC with step size control computed via “delay coefficients”.

where

- k is the time index.
- i is the subband index.
- $\mu_i(k)$ is the step size parameter for subband i.
- $\mathbf{x}_i(k) = [x_i(k)\ x_i(k-1)\ \cdots\ x_i(k-L_i-1)]^T$ is the complex vector of the reference signal at subband i.
- $\mathbf{h}_i(k) = \left[h_{0_i}\ h_{1_i}\ \cdots\ h_{(L_i-1)_i}\right]^T$ is the complex vector of the estimated impulse response at subband i.
- L_i is the length of the estimated impulse response at subband i. If L_i are the same for every subband, L_i can be set to $\frac{L}{K}$, where L is the length of the fullband estimated impulse response ($\mathbf{h}(k)$).
- $e_i(k) = d_i(k) - \mathbf{h}_i^H(k)\mathbf{x}_i(k)$ is the complex error signal at subband i.
- $d_i(k)$ is the complex far microphone signal at subband i.

The step size in each subband is computed in the following way

$$\mu_i(k) = \frac{E[|x_i(k)|^2] E\left[|\mathbf{m_i}(k)|^2\right]}{E\left[|e_i(k)|^2\right]} \tag{2}$$

where $|\mathbf{m_i}(k)|^2 = |\mathbf{g_i}(k) - \mathbf{h_i}(k)|^2$ is the system distance for the subband i. The powers of $x_i(k)$ and $e_i(k)$ are estimated via IIR filtering with different smoothing constants for rising and falling edges [7, 1]. The system distance is estimated via the method called “delay coefficients” [6, 7], given:

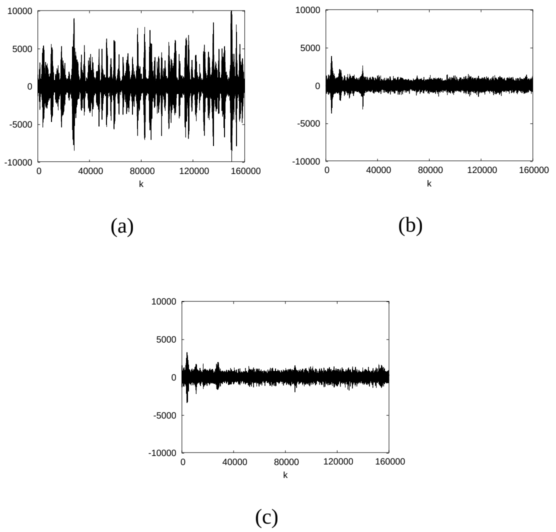
$$|\mathbf{m_i}(k)|^2 = \frac{L_i}{L_{d_i}} \sum_{j=0}^{L_{d_i}-1} |h_i(k)|^2 \tag{3}$$

where L_{d_i} is the artificial delay (in samples) introduced in each subband. In our work L_{d_i} is the same for all the subbands and is defined as: $L_{d_i} = L_d/K$, where L_d is the artificial delay introduced in the system before the AFB. In Figure 1 the system block diagram for SAEC with step size control computed via “delay coefficients” is reported.

3. COMPARISON BETWEEN FAEC AND SAEC

The most critical issue of NLMS is the low speed of convergence in presence of colored signals such as speech and in-car noise. In order to improve the convergence performance SAEC represents a possible solution. However, the convergence improvement is not evident in absence of colored background noise, due to the use of non ideal filter banks [7].

On the other hand, under noisy conditions (such as in-car background noise) the speed of convergence of SAEC increases. In fact NLMS identifies the system impulse response by minimizing the error on the full band. Then, when the error is smaller than the background noise, NLMS cannot adapt anymore. If the background noise is more concentrated in the low frequencies, then residual echo in high frequencies can not be removed this way degrading the recognition performance. Otherwise SAEC identifies the system impulse response in each subband independently, so it is able to cancel the interfering signal in the frequencies where the background noise is not present. In Figure 2 an example of a simulated far microphone signal with the signal after having applied SAEC or FAEC is reported. By inspecting Figures 2 b) and c) it seems that the acoustic echoes are well suppressed both for SAEC and FAEC, but the convergence speed in



(a)

(b)

(c)

Figure 2: a) Far microphone signal b) SAEC output signal c) FAEC output signal.

terms of normalized misalignment, defined as $\frac{\|\mathbf{h}(k)-\mathbf{g}(k)\|^2}{\|\mathbf{g}(k)\|^2}$ [10] and reported in Figure 3, shows that SAEC yields a better convergence than FAEC, because of its capability to remove echo in the subbands where background noise is not present.

Moreover SAEC can reduce the computational load in case of long impulse responses. The computational complexity for FAEC in terms of real multiplications number (N) is given mainly by the filtering and is proportional to:

$$N_{\text{FAEC}} \approx 2L \tag{4}$$

The computational complexity for SAEC is mainly given by the sum of complexity of analysis and synthesis filter banks and complexity of adaptive filtering [7], that is

$$\begin{aligned} N_{\text{SAEC}} \approx \quad & N_{\text{filter bank}} + N_{\text{adaptive filtering}} \approx \\ \approx \quad & \left(3\frac{L_p}{K} + 3\frac{M}{K}\log_2 M\right) + \left(4L\frac{M/2+1}{K^2}\right) \end{aligned} \tag{5}$$

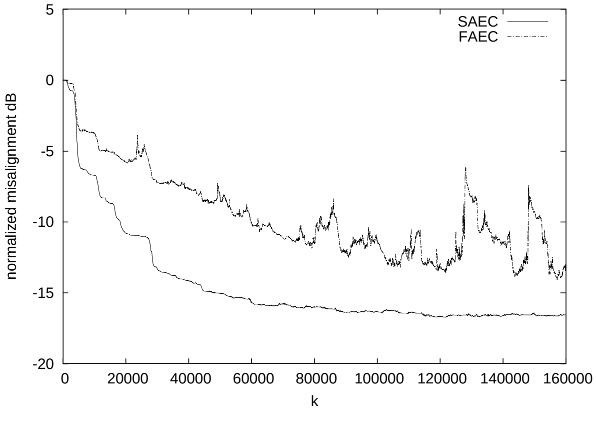


Figure 3: Comparison between the normalized misalignment of SAEC and FAEC.

Since here the system parameters are: $L = 500$, $M = 32$, $K = 28$, $L_p = 320$, we have $N_{\text{FAEC}} \approx 1000$ and $N_{\text{SAEC}} \approx 93$.
A drawback of SAEC is the introduction of a time delay which depends on the length of the prototype filter, but with an appropriate choice of it this problem is not so severe.
The protoype filter length influences the aliasing and the good quality of the reconstruction of the signal after the analysis and synthesis filter banks.
The above mentioned technique was applied as preprocessing of a speech recognition system which is described in the following section.

4. HMM RECOGNIZER

The ASR system is based on a speaker independent continuous density Hidden Markov Model (HMM)- based technology developed at ITC-irst. HMM models correspond to a set of 34 acoustic-phonetic units. All of the HMMs have a left-to-right topology and use output probability densities represented by means of 64 Gaussian components with diagonal covariance matrices. HMM training was accomplished through the standard Baum-Welch training procedure.
The front-end processing is based on the computation of 12 Mel scaled Cepstral Coefficients (MCCs) and the log-energy for each analysis frame. The analysis step is 10 ms (with a Hamming window of 20 ms) and Cepstral Mean Subtraction is applied on the entire length of the signal while the energy is normalized with respect to the maximum energy value. The resulting parameters, together with their first and second order time derivatives, are arranged into a single observation vector of 39 components.
The training is characterized by 240 speakers, 5 hours of noisy speech. Training was accomplished by using connected digit sequences.
For the test set we used clean (close-talk) speech signals extracted from the publicly available SPK Italian database [11].

5. TEST RESULTS

From the above mentioned clean database a noisy speech corpus was created artificially in the following way. The prompt speech was filtered with the measured system impulse response between the loudspeaker and the far microphone. A sample of car background noise was recorded. The user speech derived from filtering the close talk signals with the car impulse response. Then the three signals were combined together to simulate the signal picked up by the far microphone. The prompt speech and the background noise were appropriately scaled to impose different values of mean SNR and SIR. SIR is defined as the ratio between the powers of the useful signal (user speech) and the interfering signal (prompt speech), while SNR is the ratio between the powers of the useful signal and the car background noise.
Note that t_s is an inner variable of the system and could significantly affect performance, since it sets the time available to the sytem for a complete convergence.

SIR/SNR [dB]	FAEC *id*	SAEC *id*	FAEC *cd*	SAEC *cd*
20/20	1.5 %	0.81 %	5.17 %	2.84 %
20/15	1.67 %	0.72 %	6.52 %	2.69 %
20/10	2.47 %	0.69 %	7.81 %	3.5 %
20/5	3.72 %	0.86 %	12.29 %	7.11 %
15/20	2.06 %	0.86 %	5.97 %	2.98 %
15/15	2.06 %	0.86 %	7.06 %	2.84 %
15/10	2.5 %	0.86 %	8.56 %	3.58 %
15/5	3.78 %	0.92 %	13.23 %	7.3 %
10/20	2.5 %	0.86 %	7.65 %	3.59 %
10/15	2.64 %	0.86 %	8.40 %	3.03 %
10/10	3.28 %	0.94 %	9.98 %	3.75 %
10/5	4.61 %	1.00 %	14.58 %	7.26 %
5/20	4.03 %	1.78 %	10.09 %	4.73 %
5/15	4.67 %	1.72%	11.13 %	4.23 %
5/10	6 %	1.75 %	12.88 %	4.42 %
5/5	7.83 %	1.92 %	17.47 %	7.51 %

Table 1: Comparison between FAEC and SAEC for different values of SNR and SIR in terms of WER % for the isolated digits and connected digits task. t_s is set to 4 seconds.

This work addresses both isolated and connected digit recognition tasks. The number of speakers is 30 (15 males and 15 females). The total number of isolated digits is 3600, while the number of connected digit strings is 1000 (for a total number of 8000 digits).

After having applied an AEC algorithm (FAEC or SAEC) the signals are manually segmented (based on an ideal Voice Activity Detector) and fed to the recognizer. In Table 1, test results for different values of SNR and SIR are reported; t_s was set to 4 seconds. The first column shows the different possible environmental conditions, the second and third columns report on the WER % results for the isolated digits (*id*) and finally the last two columns report on the WER % results for the connected digits (*cd*). One can notice the large improvement that SAEC yields particularly in presence of high background noise and interfering signal; for a SIR = 5 dB and a SNR = 5 dB the improvement is from 7.83 % WER to 1.92 % WER in the case of isolated digit task. The performance improves from 17.47 % to 7.51 % in the case of connected digit task.

Figure 4 reports on WER % as a function of t_s in the connected digit task, given three different environmental conditions. One can notice that SAEC yields a better performance than FAEC in all of the cases.

6. CONCLUSIONS

In this work we investigated on a subband acoustic echo canceller for the introduction of a barge-in functionality in a hands-free in-car spoken dialogue system. A comparison between subband and fullband echo canceller showed the greater effectiveness of the former in terms of speed of convergence, robustness against noise and computational load. Morover SAEC yields a greater system performance in terms of WER %. Under the most adverse environmental conditions (SNR = 5 dB, SIR = 5 dB) it is able to recover 54 % of errors for the connected digit task and 57 % of errors for the isolated digit task.

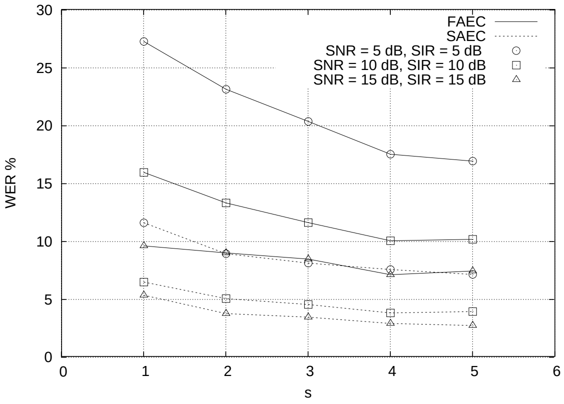


Figure 4: WER % as a function of t_s for the connected digit task and three different environmental conditions.

Finally, note that a study is needed on the possible errors introduced by a real VAD (here an ideal VAD was assumed). Future work includes the development of a real VAD integrated in the system and the analysis of its influence on the system performance.

7. REFERENCES

[1] M. Matassoni, M. Omologo, and C. Zieger, “In-car audio compensation based on nlms for hands-free speech recognition,” in *International Conference on Acoustics*, Kyoto, Japan, April 2004, pp. 2591–4594.

[2] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 2002.

[3] W. Kellermann, “Analysis and design of multirate syetms for cancellation of acoustical echoes,” in *ICASSP*, New York, 1988, pp. 25710 – 2573.

[4] A. Gilloire and M.Vetterli, “Adaptive filtering in subbands with critical sampling: Analysis, experiments and applications to acoustic echo cancellation,” *IEEE Trans. on Signal Processing*, vol. SP-40 (No.8), pp. 1862–1875, August 1992.

[5] S. Weiss, L. Lampe, and R. Stewart, “Efficient subband adaptive filtering with oversampled gdft filter banks,” in *IEEE/IEE Int. Workshop on Acoustic Echo and Noise Cancellation (IWAENC)*, London, September 1997, pp. 148–151.

[6] S. Yamamoto and S. Kitayama, “An adaptive echo canceller with variable step gain method,” *Trans. IECE Japan*, vol. E 65, Jan. 1982.

[7] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, “Acoustic echo control: An application of very-high-order adaptive filters,” *IEEE Signal Processing Magazine*, July 1999.

[8] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley, 2004.

[9] G. Schmidt, “Step-size control in subband echo cancellation systems,” in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Pocono Manor, Pennsylvania, USA, September 1999, pp. 116–119.

[10] J. Benesty and Y. Huang, *Adaptive Signal Processing*, Springer, 2003.

[11] *http://www.elda.org*.