EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

ON THE ROBUSTNESS OF THE CRITICAL-BAND ADAPTIVE FILTERING METHOD FOR MULTI-SOURCE NOISY SPEECH RECOGNITION

G. Nokas, E. Dermatas and G. Kokkinakis
Wire Communications Laboratory
Electrical & Computer Engineering Dept.
University of Patras, 26100 Patras, Greece.
Tel. +30 61 991 722, FAX: +30 61 991 855, E-mail:nokas@george.wcl2.ee.upatras.gr

ABSTRACT

In this paper we study the influence of the sub-band adaptive filtering speech enhancement method on speech recognition systems in multi-source noisy environment using a speaker and a noise reference microphone.

In extensive experiments, the recognition score of a speaker independent isolated word speech recognition system based on a continuous density HMM (CDHMM) has been measured in the presence of real life noises in various SNRs. In all experiments the results show improvement in the mean recognition score when the sub-band adaptive filtering LMS method is used in comparison to the full-band LMS method. This improvement increases when changing types of noise distort the speech signal.

1. INTRODUCTION

Multi-channel adaptive filtering is a very efficient method for speech enhancement, giving almost optimal performance even though the clean signal and the noise signal differ significantly from channel to channel and the spectral characteristics of both signals are unknown *a priori* [1]. The more input channels are available containing correlated signal components, the better the system performance is. On the other hand the speech recognition rate in noisy environment is increased by using multiple (up to four in practice) microphones [6]. In case where only two sensors are used, the multi-channel adaptive filtering techniques can be simulated by decomposing the input and the reference signal in the frequency domain (in the way the human ear does) using critical-band filters [4].

Recently, it has being shown that applying the well known LMS algorithm on the same band signals, faster convergence and more effective noise suppression can be achieved [2,5,7].

This paper compares the recognition score of a speaker independent isolated word recognition system in a multi-source noisy environment, when the critical-band LMS adaptive filtering (SB-LMS) method is used for speech enhancement with the recognition score achieved by using the full band LMS (FB-LMS) algorithm (fig. 1).

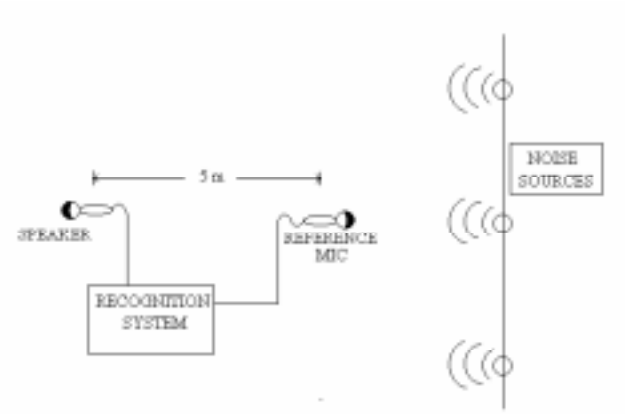

Figure 1. The multi-channel speech recognition system.

The experimental results show that the recognition rate is significantly better (up to 20%) in the first case when the SNR of the training recordings differs from that of the testing data or different types of noise were added to the testing recordings. The SB-LMS and the critical band features extraction method in conjunction with HMM parameter adaptation techniques can be used to implement speech recognition systems having acceptable recognition rates in a wide range of SNR and type of noises.

2. ADAPTIVE NOISE CANCELATION - FEATURE EXTRACTION

The noisy signals are decomposed in band-limited signals, non-linearly distributed in the frequency domain. The LMS algorithm is applied to each band to suppress the noise components. Specifically the speaker and the noise reference signals are sampled at 16 kHz, preemphasized, and decomposed in critical rectangular bands (the first 20 bands from [4], page 142) with the use of the 32 ms FFT, computed every 5ms. Fifty coefficients of linear FIR filters are estimated by the adaptive LMS algorithm for each critical band separately. The feature vector consists of the normalized log-energy of the critical-band with respect to the total frame log-energy.

3. THE SPEECH RECOGNITION SYSTEM

The experiments were carried out on a speaker independent isolated word recognition system, which is based on a whole-word CDHMM [3]. Each word model is a five states left to right CDHMM with no state skip. The output distribution probabilities are modeled by means of a Gaussian component with diagonal covariance matrix.

The segmental k-means training algorithm is used to estimate the HMM parameters from multiple feature vectors. In particular, in all experiments the maximum number of the word models was set to 1.

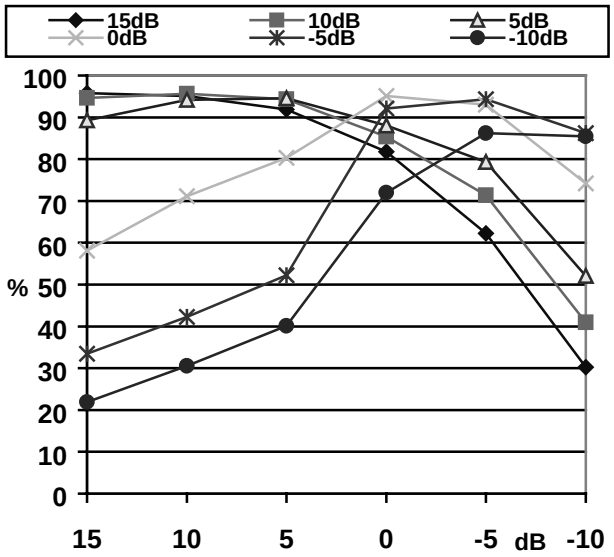

Figure 2. Recognition score (%) for the SB-LMS method versus the SNR of the testing (horizontal axis) and the training set (curves) for F16 and Car noise.

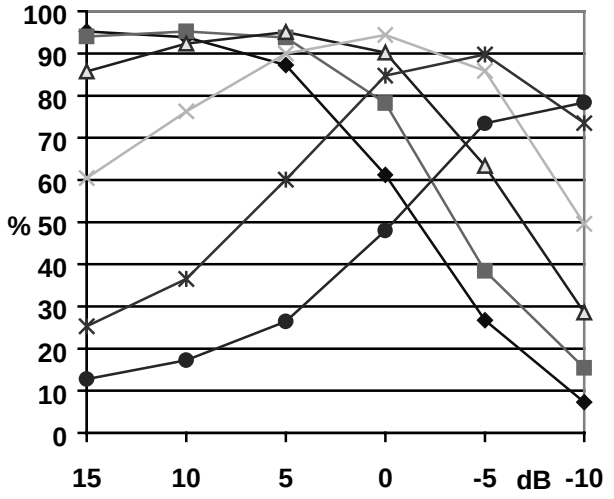

Figure 3. Recognition score (%) for the FB-LMS method versus the SNR of the testing and the training set for F16 and Car noise.

The speech database used included 15 command words and the digits of the Greek language recorded by 107 speakers in a semi-anechoic chamber. The recordings of 35 randomly selected speakers composed the training set (840 recordings) and the remaining set of 1710 recordings were used as testing set. In the recognition experiments, we used manually determined word boundaries.

4. EXPERIMENTS AND RESULTS

Three types of colored noise (taken from the NOISEX92 database) were added to the speech signal at 15, 10, 5, 0, -5, -10 dB: noise recordings under various driving conditions (Car), factory noise in a car production hall near to a plate-cutting and electrical welding equipment (Factory) and a cokpit noise under various flight conditions of a F16 fighter (F16). In addition we added two sine signals of 400 and 900 Hz respectively to the speech recordings in order to study the overall system performance in narrow band noise (Sine).

The noisy signals were created by simulation. Specifically, the speaker microphone was positioned in 5m distance from the reference noise microphone. In three experiments, different noise sources were distributed in the neighboring space in the same plane and in various distances from the microphones: (a) F16 and Car, (b) F16, Car and Factory, (c) F16, Sine, Car, Factory (fig. 1). Fig. 2 to 9 presents the experimental results.

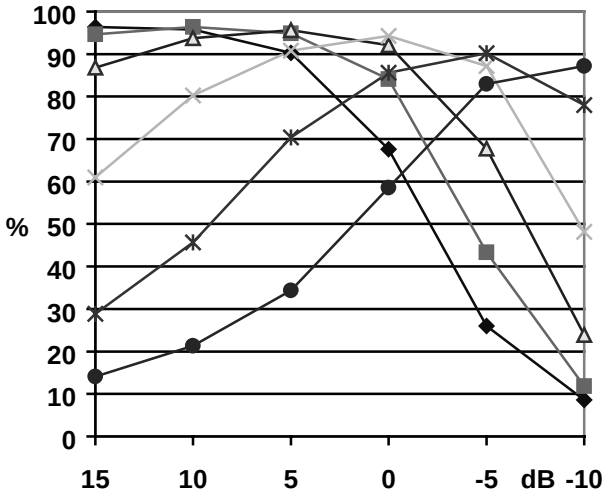

Figure 4. Recognition score (%) for the SB-LMS method versus the SNR of the testing and the training set for F16, Car and Factory noises.

The SB-LMS gives substantial better results in case where the SNR of the testing set differs from that used in the training set as shown in fig. 2 and fig. 3. The most impressive result was reached when noisy speech data at 10 dB SNR were used to train the system and the recognition rate was measured in testing data at -10dB SNR. The FB-LMS gave 11.81% while the SB-LMS improved this score to 40.99%. In general this improvement decreases when the SNR of the training recordings also decreases, as measured by the mean recognition score of the experiments carried out in different testing data SNR (table 1.).

Table 1
Mean recognition score for different training SNR

Training SNR (dB)	15	10	5	0	-5	-10
Sub - band (%)	76	80	82	78	66	56
Full - band (%)	64	70	76	76	66	49

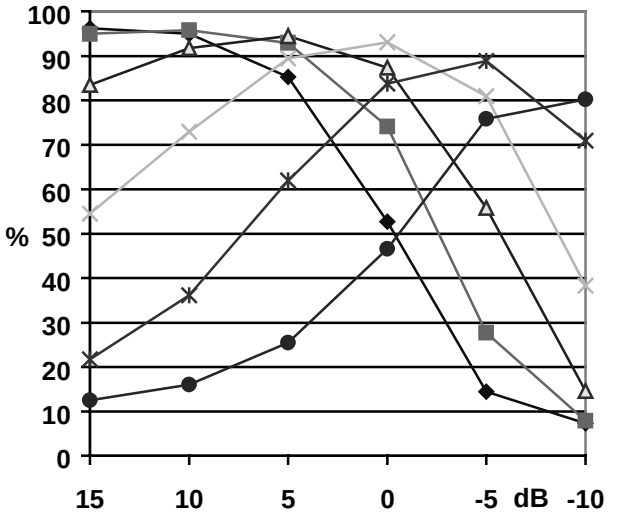

Figure 5. Recognition score (%) for the FB-LMS method versus the SNR of the testing and the training set for F16, Car and Factory noise.

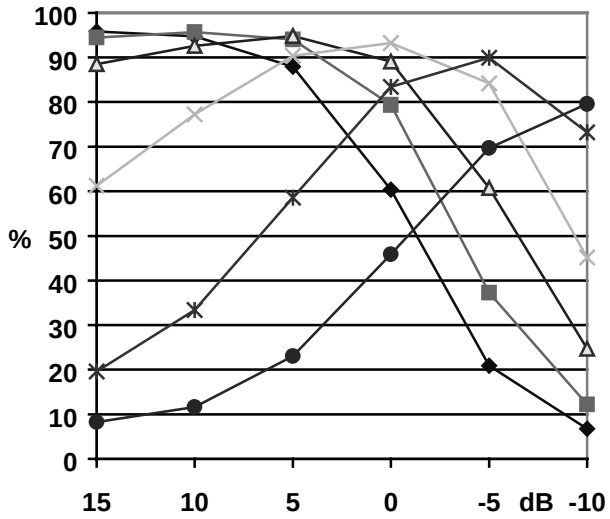

Figure 6. Recognition score (%) for the SB-LMS method versus the SNR of the testing and the training set for F16, Car, Factory and sine noise.

Table 1 gives the mean recognition rate for the experiments with the F16 and Car types of noise shown in figures 2 and 3. In all cases the SB-LMS method gives a better score, maximizing the difference from the FB-LMS method when the "clean " (high SNR) or "very noisy" speech data (-10dB) is used to train the system.

In fig. 4 and 5 the recognition rate is shown when additional narrow band and strongly non-stationary noise (factory) is used to distort the speech signal. The recognition rate decreases insignificantly in the SB-LMS method in testing data of high SNR. On the contrary, in low SNR testing data the recognition rate decreases approximately 20% (fig. 2 and 4). In the case of the FB-LMS experiments (fig. 3 and 5) the recognition rate decreases more uniformly, both in low and high SNR.

In the last set of experiments additional very narrow band stationary noise was added to the speech. Two clean sine signals of 100 and 900 Hz respectively were added to the noisy speech. This type of noise has minor influence on the recognition rate using the SB-LMS method (fig. 4 and 6 are almost identical; maximum difference of 5%). Similar phenomena were measured for the FB-LMS method (fig. 5 and 7).

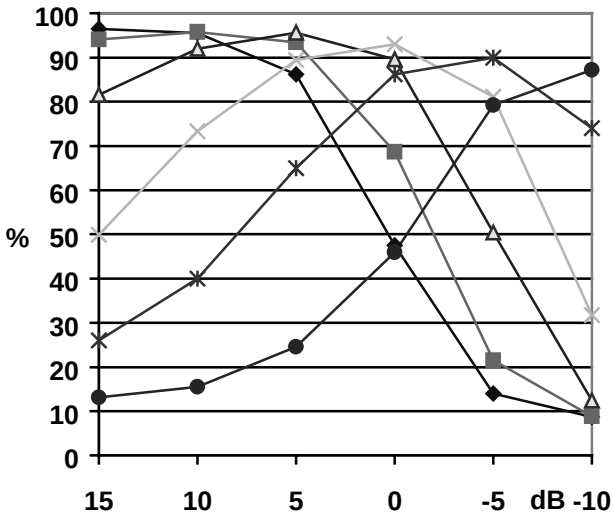

Figure 7. Recognition score (%) for the FB-LMS method versus the SNR of the testing and the training set for F16, Car, Factory and sine noise.

The influence of changing types of noise to the recognition score has been studied in two experiments. In the first experiment the speech data and all types of noises were used to train the system; noise from F16 and Car were added to testing data (fig. 8). In the second experiment (fig. 9) the role of the training and the testing data was reversed. The results show that in high SNR testing data the recognition score is greater in the case where the system is trained using all the types of noise sources. The recognition rate drops from 80% (training data of clean speech and all types of noise at -5dB, fig. 8) to 12% (training data of clean speech and F16 plus Car noise at -5dB, fig. 9) for the SB-LMS method. This phenomenon is reversed when the noise of the testing data increases. As shown in fig. 9 the best performance is achieved for the SB-LMS when the recognition system is trained using more or less clean speech data (F16 and Car noise at +10dB). The recognition rate remains high, in the range 85-94% for all testing SNR, while it decreases dramatically in the case that training includes all types of noise (fig. 8).

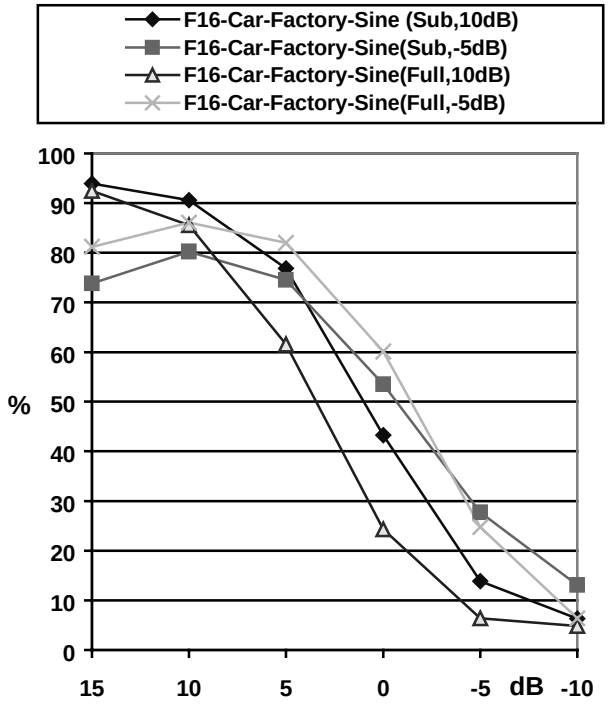

Figure 8. Recognition score (%) for the SB-LMS and FB-LMS method versus various SNR ratios of the testing data; F16 and Car noise distort the testing data at 10 and -5 dB SNR.

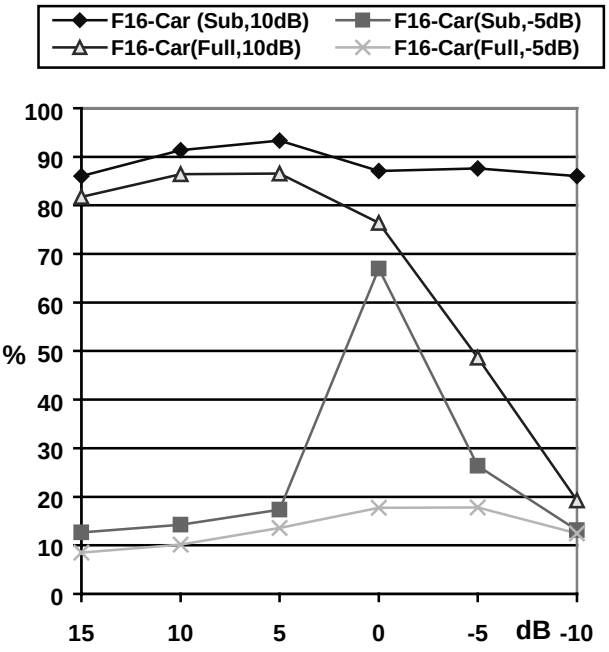

Figure 9. Recognition score (%) for the SB-LMS and FB-LMS method versus various SNR ratios of the testing data; F16, Car, Factory, sine noise distort the testing data at 10 and -5 dB SNR.

5. CONCLUSIONS

Extensive experiments have shown that the critical-band multi-channel adaptive filtering method using FIR linear filters can be employed to improve the score of isolated word recognition systems. The most important results can be summarized as follows:

a. In case the recognition system is operating in a wide area of SNR (+15,-10dB) it is better to use training data at 5dB SNR, giving a mean error rate of 82% (table 1).
b. If the recognition system is operating in a specific noise environment and in a restricted area of SNR, the most effective training data are the speech noisy data at the same SNR and type of noise.
c. The SB-LMS method gives better recognition rates than the FB-LMS in almost all the cases. This improvement is maximized when stationary narrow band noise distorts the speech signal. Though in non-stationary narrow band noise the SB-LMS method is less efficient, it remains better than the FB-LMS.
d. In case the recognition system is operating in various noisy environments (changing types of noise) it is better to train the system using high SNR speech data containing the most frequent types of noise. In the recognition phase the SB-LMS can be used efficiently to minimize the influence of the unknown types of noise.

6. REFERENCES

[1] E. Ferrara E. and B. Widrow, "Multichannel Adaptive Filtering for Signal Enhancement", IEEE Tr. on Acoustics, Speech, and Signal Processing, Vol. 29, No 3, pp. 766-770, 1981.

[2] D. Darlington and D. Campbell, "Sub-Band Adaptive Filtering Applied To Speech Enhancement", ICLSP, pp. 921-924, 1996.

[3] E. Dermatas and G. Kokkinakis, "Algorithm for Clustering Continuous Density HMM by Recognition Error", IEEE Tr. on Speech and Audio Proc., Vol. 4, No 3, pp. 231-234, 1996.

[4] E. Zwicker E. and H. Fasl, "Psychoacoustics: Facts and Models", Springer Verlang, Berlin Heidelberg, 1990.

[5] A. Sugiyama and F. Landais, "A New Adaptive Intersubband Tap-assignment Algorithm for Sub-band Adaptive Filters", ICASSP, pp. 3051-3054, 1995.

[6] D. Giuliani, M. Omologo, and P. Svaizer, "Experiments of Speech Recognition in a Noisy and Reverberant Environment using a Microphone Array and HMM Adaptation", ICLSP, pp. 1329-1332, 1996.

[7] J. Shynk, "Frequency-Domain and Multirate Adaptive Filtering", IEEE SP Magazine, pp. 14-37, Jan. 1992.