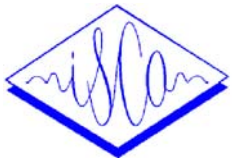


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEECH RECOGNITION IN NOISY REVERBERANT ROOMS USING A FREQUENCY DOMAIN BLIND DECONVOLUTION METHOD

G. Nokas and E. Dermatas

Wire Communications Laboratory, Electrical & Computer Engineering Department.,
University of Patras, 26100 Patras, Hellas. Tel. +30 61 991722, FAX: +30 61 991855,
E-mail:nokas@george.wcl2.ee.upatras.gr

10.21437/Eurospeech.1999-631

ABSTRACT

The aim of this paper is to present and evaluate two adaptive speech enhancement methods in the frequency domain by measuring the recognition rate of a speaker-independent word recognition system of isolated words.

In a hands-free speech recognition experiment, a factory noise source and a speaker are the acoustic sources in a real room environment. A close-talking microphone is positioned near the noise source while the primary omni-directional microphone captures the convoluted speech and noise signals. Adaptive noise cancellation reduces the presence of noise in the primary microphone followed by a blind deconvolution algorithm used to minimize reverberations.

Experimental results showed that the proposed speech enhancement methods increase 2.4 times the recognition rate for a vocabulary of 21 words of the Greek language, but the achieved recognition rate of 16% is inapplicable for commercial applications.

INTRODUCTION

Speech recognition in noisy reverberant rooms requires robust speech enhancement methods. In case of hands-free speech recognition applications, the microphone signal is the convolution of the original signal, the impulse response of the transducer, and the impulse response of the surrounding environment by assuming linear propagation of acoustic signals through the air.

The Hidden-Markov-Model (HMM) classifier achieves best performance when training and testing features are detected in similar operation conditions. In practice, clean speech features are used to train the HMM and, as the operation environment varies, the divergence from the above rule decreases significantly the recognition accuracy.

To overcome this problem, three types of solutions have been proposed; the use of robust features, the implementation of adaptation methods for the speech models, or filtering the noisy speech prior to recognition.

Stochastic gradient algorithms applied to acoustic channel models are one of the most promising approaches to the noise reduction problem using filtering techniques. The reconstruction of the clean speech in noisy reverberant rooms requires adaptive filtering of noise and deconvolution of the room transfer function. The well known least-mean-square algorithm (LMS) adapts the parameters of linear noise reduction filters to any environment where multilple noise sources are present using an array of microphones. In this case the LMS implementation requires the exact knowledge of the noise signals.

This paper studies the case where a noise source and a speaker are the acoustic sources in a real room environment. A close-talking microphone is positioned near the noise source while an omni-directional microphone captures the convoluted speech and noise signals [1]. In this environment a speech recognizer of isolated words is evaluated employing two speech enhancement frontend processors. Taking into account that in practice the statistics of the signals are unknown and time variant, a linear adaptive noise cancellation (ANC) processor is used to reduce the presence of noise. The enhanced signal, in case of perfect adaptation, consists of the clean speech convoluted with the impulse response of the transfer function between the distant microphone and the speaker. Blind deconvolution techniques can achieve channel equalization without using training signals, by bypassing the estimation of channel length and introducing only reduced complexity. In our implementation a Bussgang-like algorithm [7,9] is used to reach the clean speech, matching the speech recognizer features used in the training and the operation mode. Both ANC and blind deconvolution (BD) algorithms are applied in the frequency domain eliminating the overall complexity of the frontend and the feature extraction processors.

In the following two sections, a detailed description of the proposed speech enhancement methods are given. The speech recognition experiments and the experimental conditions are presented in section 3. The experimental results and a short discussion conclude this work.

1. ADAPTIVE FILTERING IN THE FREQUENCY DOMAIN

The concept of noise suppression by using frequency domain adaptive filters (FDAF) in reverberant rooms can be described as follows: The signal n(t) of a non-stationary noise source is captured by a close-talking microphone positioned near the source. We assume that the environmental interference in the reference signal n(t) can be neglected. The primary sensor x(t), an omni-directional microphone, receives the speech and the noise signal after reverberations which consist of direct

and multipath reflections between the two signal sources and the receiver (Fig.3).

The block diagram of the noise cancellation method in the frequency domain (first presented by Dentino et. al [10]) is shown in Fig.1.

Initially, the reference signal n(t) is delayed by the signal propagation time through the direct acoustic path between the noise source and the primary microphone. The time delay can be set externally in case of known topology, but in practice an accurate estimation can be achieved at the time value corresponding to the maximum of the cross-power spectrum phase of the primary and the reference signal.

$$T = \text{argmax}\ \ \text{invFFT}[\, N(f)\, X^H(f)]$$

where, X(f), N(f) are the FFT of the primary and the reference signal respectively and, $X^H(f)$ is the conjugate of X(f).

The synchronization process of the noise signals between the primary and the reference microphone maximizes the speech enhancement capabilities of the proposed method.

The signal of the primary sensor is decomposed into two convoluted signals:

$$x(t) = Fs(t) * s(t) + Fn(t) * n(t),$$

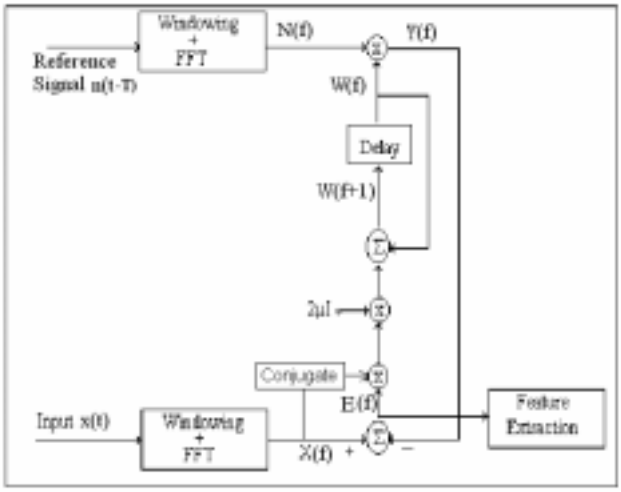
where, Fs(t) and Fn(t) are the room impulse responses of the speech and noise sources respectively, and s(t) is the clean speech signal.

The primary x(t) and the delayed reference signal n(t-T) are windowed and transformed using FFT.

The LMS minimization of the following error function:

$$E(f) = X(f) - N(f)W(f) \qquad (1)$$

(W(f) are the complex coefficients of the linear filter) leads to the following tap-weight adaptation applied to each FFT frame:



$$W(f) \leftarrow W(f) + 2\mu X^H(f)E(f) \qquad (2)$$

Figure 1. Frequency domain adaptive filtering and feature extraction.

At the global minimum, the error signal becomes the FFT of the clean speech signal arriving in the primary sensor:

$$E(f) = FFT(Fs(t) * s(t))$$

2. ADAPTIVE BLIND DECONVOLUTION

The problem of estimating the clean speech in a reverberant environment with the knowledge of the clean speech statistics, is known as blind deconvolution or unsupervised deconvolution. We can identify two families of blind deconvolution algorithms, those based on Higher Order statistics and those based on Cyclostationary statistics. In this study we introduce an adaptive blind deconvolution scheme implemented in the frequency domain which has many similarities with the Treichler's, well known Constant Modulus Algorithm (CMA) [7].

2.1 Blind Deconvolution in the time domain

In the time domain, the most widely used algorithm for blind deconvolution, is the CMA which is part of the Bussgang family algorithms. In this approach the only given information is the convoluted input data x(t) and higher order statistics of the original signal. The algorithm adjusts the coefficients of an FIR filter in order to minimize the cost function:

$$J = E[(|y(t)|^2 - R)^2] \qquad (3)$$

where, y(t) is the filter output and R is the constant modulus:

$$R = \frac{E[|s(t)|^4]}{E[|s(t)|^2]}$$

(s(t) is the clean signal, and E[] the expectation operator).

The objective in this adaptive filtering is to restore y(t) in a form which, on the overage, has a constant instantaneous modulus based on the "a priori" statistics of the clean non-reverberant signal.

2.2 Blind Deconvolution in the frequency domain

In our implementation, we are interested in the deconvoluted energy spectrum at each speech frame, in order to extract the speech recognizer features. Thus, FFT and energy estimation in each critical band (in the way the human ear does using the first 20 bands from [3], page142) is performed on the basis of overlapped-windowed frames. The energy vector is normalized to the unity, in order to achieve equalization of signals with different amplitude levels.

The normalized energy of each critical band Y(m) is multiplied by a real coefficient W'(m). The coefficients W'(m) are adjusted in order to minimize the following cost function:

$$C = E[(Y(m) - R(m))^2] \quad (4)$$

where R(m) is a reference spectrum defined by the mean normalized energy spectrum of the clean speech data.

Equations (3) and (4) are similar, since they try to bound the signal energy in given limits, except that in equation (4) the error is a linear function of the filter output. The LMS algorithm yields the optimum coefficients of a linear filter minimizing the error function of equation (4).

The above filtering method, since it is not equivalent to a circular convolution in the time domain, gives a good approximation of the time domain optimum solution requiring minimal computational effort. Mokbel et al [8] report improvements of speech recognition rates after equalization of the telephone lines effects using this approximation.

The FDAF and BD algorithms are applied in the frequency domain, thus a compact, easy to implement and fast procedure results. The block diagram of both FDAF and BC algorithms is presented in fig. 2, where the two microphone signals are processed in the frequency domain and the output is the feature vector of the recognition system.

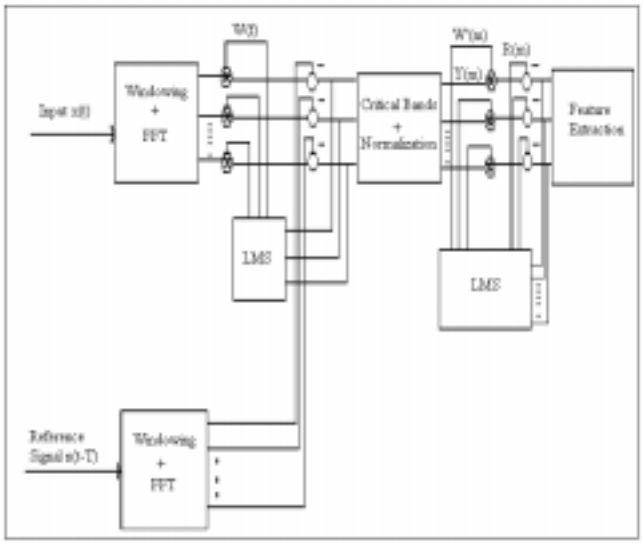


Figure 2. Adaptive LMS noise cancelation and blind deconvolution in the frequency domain.

3. EXPERIMENTS

The speech database used for the method's evaluation consists of 11 command words and the ten digits of the Greek language recorded by 76 speakers in an anechoic chamber (clean recordings). The recordings of 35 randomly selected speakers composed the training set and the remaining recordings were used for testing purposes. In the recognition experiments, we used manually determined word boundaries. A loudspeaker in a reverberant room (our lab) reproduced the speech recordings. Another loudspeaker was used to produce real-life non-stationary noise (the factory noise of the NOISEX92 [6] database). Two-channel recordings of each word were created with a close-talking cardioid microphone near the noise loudspeaker and one distant microphone in the front of the speaker loudspeaker as shown in fig. 3. Acquisitions were carried out synchronously for both input channels and were sampled with 16kHz and 16bit accuracy. Acquisitions were also carried out without the noise source, producing a noiseless reverberant database.

In our experiments the distance between the two microphones was set to 3 meters. A close-talking microphone was placed at 10 centimeters from the noise loudspeaker and the distance between the primary microphone and the speaker was 1 meter.

Interference in the reference noise signal decreases significantly the noise reduction capabilities of the FDAF methods in reverberant environment [4]. Thus, the reference microphone was positioned very close to the noise source. In case of known positions of the microphones and the acoustic sources, the time delay of the noise signal in the primary microphone can be estimated from direct path differences [5].

The experiments were carried out on a speaker independent isolated word recognition system, which is based on a whole word CDHHM [2]. Each word model is five states left to right CDHMM with no state skip. The feature vector consists of the normalized log-energy of the critical-bands with respect to the total log-energy of the frame. The output distribution probabilities are modeled by means of a Gaussian component with diagonal covariance matrix. The segmental k-means training algorithm was used to estimate the HMM parameters from multiple feature vectors.

4. EXPERIMENTAL RESULTS

In the first experiment the FDAF noise canceling method was evaluated. Initially, the isolated word recognition system was trained using noiseless reverberant recordings and tested in real-life noise (recognition rate of 34%). The mean SNR in the primary microphone was approximately 10db without noise canceling (energy in the critical bands). After incorporating the FDAF method to the speech recognition system, the same experiments were carried out (recognition rate of 51%). For comparison purposes, the recognition rate in case of training the recognizer using noisy reverberant recordings was also measured and provided an accuracy of 89%.

In the second experiment the BD method in the frequency domain was evaluated in a computer room environment (fan noise is present in addition to the artificial noise). Initially, the clean speech data was used to train the speech recognition system. Speech recordings reproduced in the noiseless reverberant environment were used to measure the recognition rate (44%). In case of removing the reverberation effects from the testing recordings using the BD method, the recognition rate increased by 11%. With matched conditions (noiseless reverberant recordings used both for training and testing) the best performance was achieved (92%).

Finally, we evaluated the combined method applying both FDAF and the BD method. The speech recognition system was trained using clean recordings and tested in a real-life noisy room environment. The recognition rate without any

speech enhancement method (SNR ≅ 10db) was 6.7%. In the case of FDAF method the rate increased to 10.5% and, in the combined method (both FDAF and BD were applied in the frequency domain) the same experiments reached 16% (table 1).

Table 1. Recognition rate of the three experiments.

1st Experiment		
Noiseless reverberant training, Noisy reverberant testing	Noiseless reverberant training, Noisy reverberant testing using FDAF	Noisy reverberant training, Noisy reverberant testing
34 %	51 %	89 %

2nd Experiment		
Clean training, Noiseless reverberant testing	Clean training, Noiseless reverberant testing using BD	Noiseless reverberant training, Noiseless reverberant testing
44 %	55 %	92 %

3rd Experiment		
Clean training, Noisy reverberant testing	Clean training, Noisy reverberant testing using FDAF	Clean training, Noisy reverberant testing using FDAF and BD
6,7 %	10,5 %	16 %

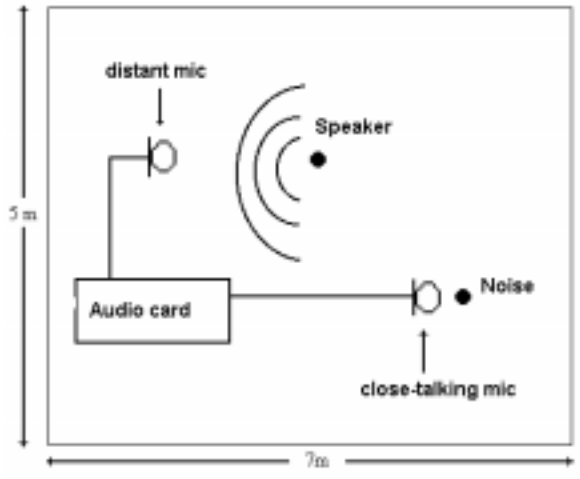


Figure 3. Experimental set up.

The experimental results showed enormous degradation of the recognition score in the case of clean training and operation in the simulated noisy reverberant environment.

The recognition scores increased by 55% in case of reducing the noise components of the primary microphone using the FDAF method. The BD method increases 25% the recognition rate in a noiseless reverberant environment. In the last experiment, the recognition rate increased by 138% in case of training the recognizer using the clean recordings and performing the evaluation experiments in a real-life noisy room. Additional experiments showed that significant improvement could be achieved in case of matching the training and operation environment.

5. CONCLUSION

In this paper we presented and evaluated adaptive noise cancellation and blind deconvolution for speech recognition systems operating in real-life noisy rooms. The implementation of these methods in the frequency domain improves significantly the recognition rate but this still insufficient for commercial applications. It is shown that in case of training the HMM parameters using clean recordings, the speaker-independent isolated word recognition rate after speech enhancement is limited to 16% for 21 words of the Greek language.

The main obstacle of implementing efficient adaptive filtering methods for speech enhancement is the statistical nature of the speech and the real-life noise that are strongly non-stationary signals.

REFERENCES

[1] S. Boll, and D. Pulsipher, "Suppression of Acoustic Noise in Speech Using Two Microphone Adaprive Noise Cancellation", IEEE Tr. ASSP-28, pp 752-753, 1980.

[2] E. Dermatas and G. Kokkinakis, "Algorithm for Clustering Continuous Density HMM by Recognition Error", IEEE Tr. on Speech and Audio Proc., 4(3), pp 231-234, 1996.

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

[4] Mean-Hoa Lu and Peter M. Clarkson, "The performance of adaptive noise cancellation systems in reverberant rooms", JASA 93(2), pp. 1122-1135, Feb. 1993

[5] Francis Reed and Paul L. Feintuch, " A comparison of LMS adaptive cancellers implemented in the frequency domain and the time domain" IEEE Trans. On circuits and systems, 28(6), pp. 610-615, June 1981

[6] A.P Varga et al. "The NOISEX-92 study on the effect of additive noise on automatic speech recognition,", Technical Report, DRA Speech Research Unit, 1992.

[7] John R. Treichler and Brian G. Agee: "A New Approach to Multipath Correction of Constant Modulus Signals" IEEE Trans. on ASSP, 31(2), pp. 459-471, April 1983.

[8] Chafic Mokbel, D. Jouvet, Jean Monne: "Deconvolution of telephone line effects for speech recognition" Speech Communication, 19(3), Sep. 1996

[9] Haykin S. "Adaptive filtering Theory", Prentice Hall.

[10]Dentino, M., McCool, J, M., Widrow, B. "Adaptive filtering in the frequency domain", Proc. IEEE. 66(12), pp. 1658-1659, Dec 1978