IST2005
INTERNATIONAL SYMPOSIUM ON TELECOMMUNICATIONS
September 10-12, 2005 Shiraz, Iran

Adaptive Time Domain Signal Estimation for Multi-Microphone Speech Enhancement

Davood Shamsi*, Massoud Babaie-Zadeh*
* Advanced Communications Research Institute (ACRI), Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
Emails: david@ee.sharif.edu, mbzadeh@yahoo.com

Abstract

In this paper, the main ideas of a new method for multi-microphone speech enhancement are presented. Combining previously known multi-channel speech enhancement methods with Maximum A Posteriori (MAP) estimation concepts, and temporally whitenning the output noises, we obtain a new algorithm called Adaptive Time-domain Signal Estimation for Multi-Microphone (ATSEM). The simulations (using real-world recorded signals) emphasize on the quality of the proposed method.

Keywords: Microphone Array, Multi-Channel signal processing, speech processing, GSC.

Introduction

In many speech communication systems, such as hands-free mobile telephony, hearing aids and voice-controlled systems, it is important to obtain a clear speech signal from speech signals which are often corrupted by a considerable amount of acoustic background noise. The speech enhancement methods are divided into two general groups: a) single-microphone methods (as spectral subtraction [1], Kalman filtering [2], and signal subspace-based techniques [3], [4]), and b) multi-microphone (multi-channel) methods.

Since the speech signal and the acoustic noise signal have the same frequency-band, in practice these methods can not reduce background noise without introducing noticeable artifacts (e.g. musical noise) or speech distortion. In fact, single microphone speech enhancement algorithms suffer from two major noises [10]: diffused noise and point-wise noise. The source of diffused noise is spread over the scanning area, while the source of point-wise noise is located in certain position.

In point-wise noise, when the speech and noise sources are physically located at different positions, this spatial diversity can be exploited by using a microphone array, such that both the spectral and the spatial characteristics of the signal sources can be used. In multi-channel speech enhancement, there are two approaches for noise reduction: using fixed beamformer and using adaptive beamformer [5]. A fixed delay-and-sum (DS) beamformer spatially aligns the microphone signals to the direction of the speech source. A well-known adaptive implementation of this beamformer is the Generalized Sidelobe Canceller (GSC) [6], which consists of a fixed beamformer, creating the so-called speech reference signal; a blocking matrix, creating the so-called noise reference signal; and a multi-channel adaptive filter which eliminates the noise components in the speech reference signal which are correlated with the noise reference signals.

In this paper, we introduce Adaptive Time-domain Signal Estimation for Multi-Microphone (ATSEM) noise reduction technique.

ATSEM uses previous data to estimate probability density function (PDF) of incoming signals and speech signal. Then, by a maximum *a posteriori probability (MAP) estimation*, it estimates pure speech signal in noisy environment. Simulation results emphasize on the better performance of ATSEM compared with previous methods such as GSC.

1. Maximum A Posteriori (MAP) method

In this section, time domain signal estimation for multi-channel noise reduction is introduced. Let $y_1[k], y_2[k]..., y_M[k]$ be M signals that are corrupted by different noises, *i.e.*

$$y_i[k] = s[k] + n_i[k] \tag{1}$$

where $s[k]$ is the desired signal which is common in all input signals and $n_i[k]$ is the noise component at the i'th input. By considering noise reduction problem as an estimation problem, desired signal (unknown parameter) should be estimated based on noisy input signals $y_i[k]$ $i = 1...M$ (observed data). If estimated signal showed by $\hat{s}$ then:

$$\hat{s} = g\left(y_1[k], y_2[k]..., y_M[k]\right) \qquad (2)$$

where g is an estimator function which estimates $\hat{s}$ base on observed data $y_i[k]$ $i = 1...M$.

Estimation theory offers lots of estimation methods to estimate pure signal form noisy input signals, such as MLE (Maximum Likelihood Estimation), LSE (Least Square Estimation) [9]. By knowing a priori probability density function of pure signal and noise component Bayesian estimation may be used to estimate pure signal.

$$\hat{s} = \arg\max_{s}\left[P(y_1[k], y_2[k]..., y_M[k] \mid s) f(s)\right] \qquad (3)$$

where $f(s)$ is the PDF of the pure signal and $P(y_1[k], y_2[k]..., y_M[k] \mid s)$ is the PDF of observation $y_1[k], y_2[k]..., y_M[k]$ where pure signal is given. If noise component of signals assumed to be independent then

$$P(y_1[k], y_2[k]..., y_M[k] \mid s) =$$

$$P(y_1[k] \mid s) P(y_2[k] \mid s) ... P(y_M[k] \mid s) \qquad (4)$$

Now we make three more assumptions:

a. Noise component in each channel is Gaussian white zero mean and specific variance.

b. Noise component in each channel is white.

c. Innovation process of speech is Gaussian.

The second assumption says that previous samples of noise hove no effect on current sample. Therefore, we do not lose any information if we use only the current sample of noise for estimation.

From the first assumption:

$$P(y_i[k] \mid s) = \frac{1}{\sqrt{2\pi\sigma_i^{\,2}}} e^{\frac{(y-s)^2}{2\sigma_s^{\,2}}} \qquad (5)$$

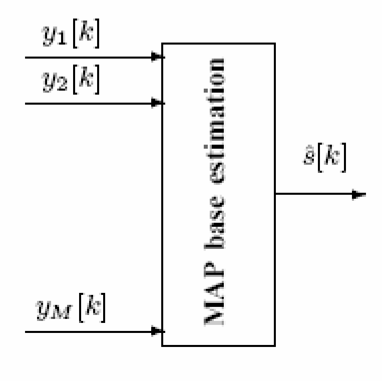

Fig. 1 MAP estimation for noise reduction

where $\sigma_i^{\,2}$ is the variance of the noise of the i'th channel.

By the third assumption, if the speech sample at the current time is predicted to be s', the PDF of speech can be written:

$$f(s) = \frac{1}{\sqrt{2\pi\sigma_s^{\,2}}} e^{\frac{\left(s-s'\right)^2}{2\sigma_s^{\,2}}} \qquad (6)$$

where σ_s^2 is the prediction error variance.

By substitution (6) and (5) in (4) and some computation, $\hat{s}$ can be easily obtain:

$$\hat{s} = \frac{\sum_{i=1}^{M} \frac{y_i[k]}{\sigma_i^2} + \frac{s'}{\sigma_s^2}}{\sum_{i=1}^{M} \frac{1}{\sigma_i^2} + \frac{1}{\sigma_s^2}} \qquad (7)$$

and finally this estimation is used as de-noised signal.

2. ATSEM Noise Reduction Method

In MAP method three assumptions were made and if we want to use this method for noise reduction, this assumption should be met. Complicated structure of ATSEM, which can be implemented easily, tries to do this.

Figure 2 shows the structure of ATSEM, which consists of six main parts. Outputs of microphone array are input of ATSEM (in figure 1 inputs of ATSEM are shown by $y_1[k]$, $y_2[k]$, ... $y_M[k]$).

First of all a fixed beamformer is used to synchronize input signals, then Blocking Matrix produces noise references for *Multi channel Adaptive Noise Canceller* (MANC), after that MANC by

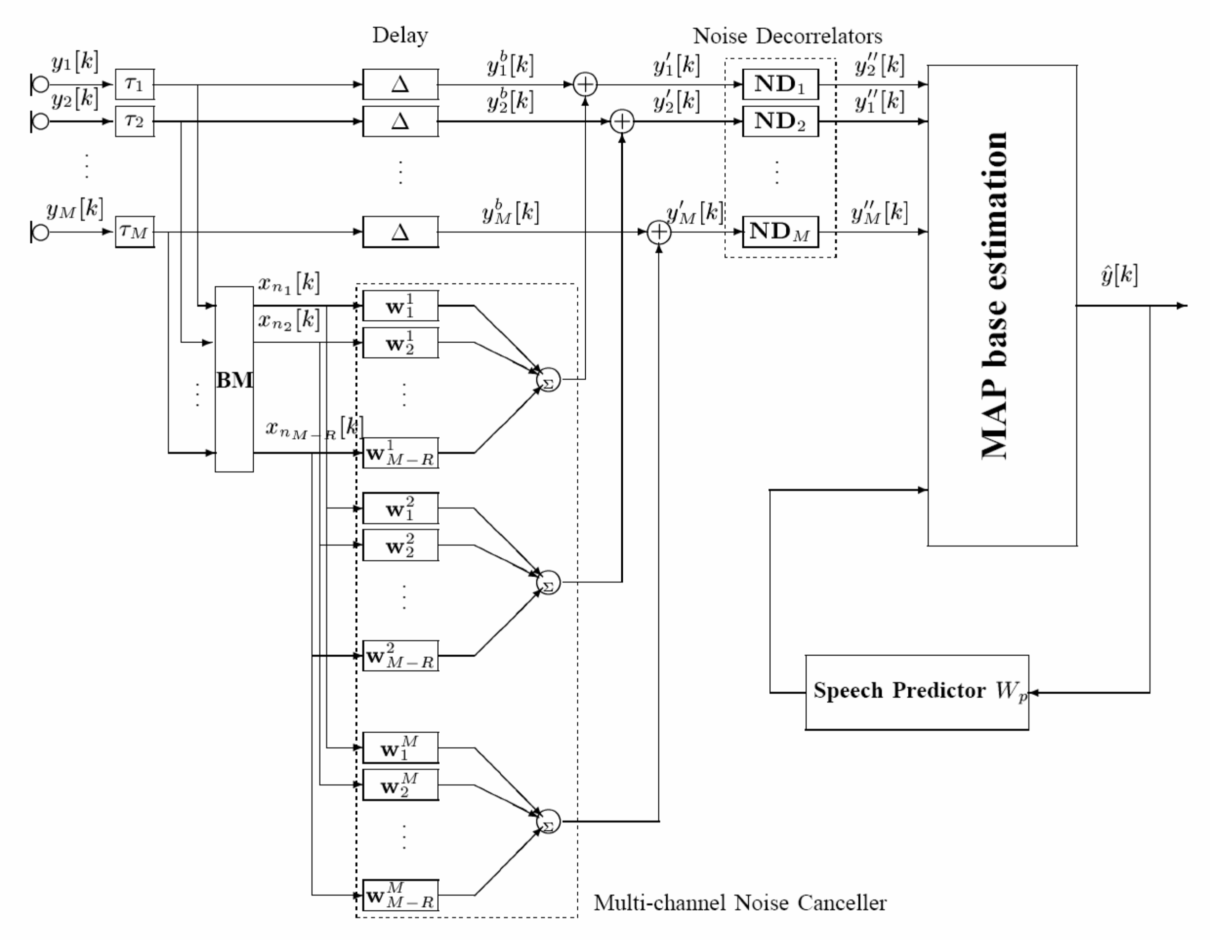

Fig. 2 Structure of ATSEM

means of noise references removes common noise between different channels, next *Noise Decorrelators* produce innovation process of incoming noise, and finally a MAP estimation is used to introduce clear signal.

2.1. Fixed Beamformer

Fixed Beamformer (FB) consists of M time delay steering elements τ_1, τ_2, … τ_M , which are used to point the array into the desired direction. In other word, these delay elements compensate received delay in different microphones and after FB speech signals have no delay respect to each other.

2.2. Blocking Matrix (BM)

Blocking Matrix is used to eliminate signal component of the microphone output and produces noise reference base on spatial diversity of noise and speech. BM first introduced by J. Griffiths [6] and currently there are a lot of algorithms based on his original idea [7]. In this algorithm a simple differentiation is used as blocking matrix. Assume the outputs of FB are $y_1^a[k]$, $y_2^a[k]$, … $y_M^a[k]$ then it can be written:

$$y_i^a[k] = s[k] + n_i^a[k] \tag{8}$$

where $s[k]$ is the pure speech and $n_i^a[k]$ is the noise component in the i'th microphone. Now noise references can easily be obtained by differentiating i'th input from $i-1$ 'th input.

$$x_{n_i}[k] = y_i^a[k] - y_{i-1}^a[k] = n_i^a[k] - n_{i-1}^a[k] \tag{9}$$

$x_{n_i}[k]$ do not contain the speech component and therefore can be used as a noise reference.

2.3. Multi-channel Adaptive Noise Canceller

Multi-channel Adaptive Noise Canceller (MANC) uses the noise references which produced by blocking matrix to remove correlation between noise in microphone outputs. In other word, MANC is used to produce channels with independent noise components, which was the main assumption in the previous section. In practice MANC is implemented by RLS or LMS updating algorithm [9].

2.4. Noise Decorrelator (ND)

In preceding section we assumed noise components are white, while in practice they are not. Noise Decorrelators (ND) are used to solve this problem. Figure 3 the shows structure of ND, which uses previous sample of noise to predict current noise sample. This can be implemented with RLS or LMS [9]. Previous samples of noise are obtained by differentiation previous noisy signal from estimated signal *i.e.*

$$n_i^{'}[m] = y_i^{'}[m] - \hat{y}[m], m < k \qquad (10)$$

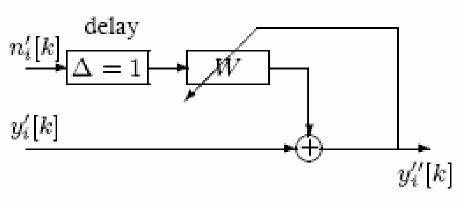

Fig. 3 Structure of Noise Decorrelator

2.5. Speech Predictor

The last part of ATSEM, which will be explained, is speech predictor. We saw in the previous section that speech prediction is needed to estimated pure signal since a speech predictor is used to predict speech, which can be implemented in various methods *i.e.* LPC. Again in speech predictor, previous estimated signals are used as previous samples of speech.

Although, ATSEM seems complicated, it can be implemented by some LMS updating algorithm. Since this algorithm consists of $2M+1$ LMS, the complexity of ATSEM is $O(2ML+1)$ where M is the number of microphones and L is the length of filters.

3. Simulation Results

This algorithm is tested in a real-world noisy environment by using three microphones in presence of diffused and point-wise noise. We have used sound file, which have been recorded in ESAT-SISTA by Simon Doclo [11]. Point-wise noise is speech-like noise from NOISEX-92, which is database of recording of various noises. Microphones were located in line array and the distance between two adjacent microphones is 5cm. The speech source is located at 1.3m from the center of the microphone array at an angle of 56 degree. In this condition output of microphones are recorded and diffused noise simulated by adding independent noise to recorded sound. GSC and ATSEM are applied to this recorded sound. Simulation shows that ATSEM results in 17.5 dB increase in SNR (Signal to Noise Ratio) as apposed to 15 dB in the GSC.

Figure 4 is one of the recorded signals and Fig. 5 shows output of ATSEM algorithm when applied to recorded signals.

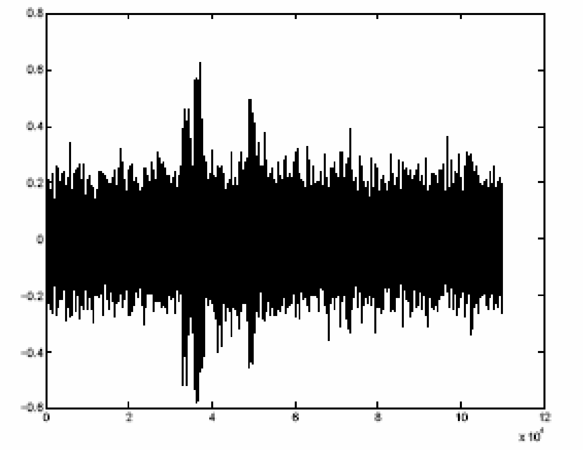

Fig. 4 A recorded signal, which is corrupted by noise

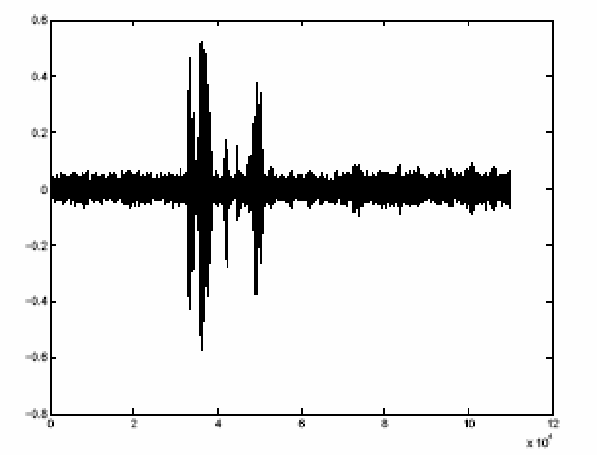

Fig. 5 Output of ATSEM algorithm

Conclusion

In this paper, we proposed the a multi-channel speech enhancement method called ATSEM, which results in better SNR compared with previously known methods. The method is tested in real-world situation.

The main drawback of the method is the presence of a set of parameters, and their fine tuning is somewhat tricky.

References

[1] Y. Ephraim and D. Malah, “Speech enhancemnt using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. on Acoust., Speech, Signal Processing,* vol. 33, no. 2, pp. 443–445, Apr. 1985.

[2] D. Burshtein S. Gannot and E. Weinstein, “Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, July 1998.

[3] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4,pp. 251–266, July 1995.

[4] S. D. Hansen S. H. Jensen, P. C. Hansen and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated qsvd," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 6, pp. 439–448, Nov 1995.

[5] B. D. Van Veen and Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[6] L. J. Griffiths and C. W. Jim, "An alternative approach to linear constrained adaptive beamforming," *IEEE Trans. Antennas Propagat*, vol. AP-30, no. 1,pp. 27–34, 1982.

[7] O. Hoshuyama A. Sugiyama A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2677–2684, October 1999.

[8] Simon Haykin, Adaptive Filter Theory, Prentice-Hall, 1996.

[9] Steven M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, 1993.

[10] Sharon Gannot, and Israel Cohen, *IEEE trans. On speech and audio processing*, VOL. 12, NO. 6, NOVEMBER 2004 561

[11] http://www.esat.kuleuvien.ac.be/~doclo/SA00061 /audio.html