Available online at www.sciencedirect.com
ScienceDirect
Applied Acoustics 69 (2008) 751–754
applied acoustics
www.elsevier.com/locate/apacoust

Technical Note

Time varying forgetting factor for the noise estimation in multi-channel noise reduction

Gibak Kim *,[1], Nam Ik Cho [1]

School of Electrical Engineering, Seoul National University, Seoul 151-744, Republic of Korea

Received 30 November 2006; received in revised form 21 February 2007; accepted 23 March 2007
Available online 23 May 2007

Abstract

We introduce a time varying forgetting factor for the noise estimation in multi-channel noise reduction. In conventional multi-channel noise reduction system, the noise statistics are estimated during noise-only periods and kept fixed during speech-present periods. For deciding whether the current frame is speech-present period or not, a voice activity detector (VAD) is usually used. Instead of this conventional scheme that needs an explicit VAD, we adopt a time varying forgetting factor which is parameterized from the normalized cross correlation (NCC). The simulation results show that multi-channel noise reduction using the proposed method yields better performance in several typical objective measures.

Keywords: Noise estimation; Multi-channel filtering; Speech enhancement

1. Introduction

Noise reduction is one of the most important elements of speech communication and voice recognition systems. Particularly in distant or hands-free speech acquisition, the system performance is severely degraded due to ambient noises. Hence, there have been many researches on noise suppression, and diverse techniques have been developed for a single or multiple microphones. Multi-microphone systems exhibit better performance than the single-microphone system since they are able to take advantage of spatial information of signals. Specifically, fixed beamforming, adaptive beamforming, and multi-channel Wiener filtering have been proposed for the multi-microphone based noise reduction system. More recently, multi-channel filtering based on GSVD (generalized singular value decomposition) or QR decomposition has been introduced, which is shown to have better performance than the standard beamforming techniques [1,2]. We focus the QR decomposition based noise reduction scheme that is computationally more efficient than the GSVD based optimal filtering [2].

Estimation of noise statistics is required in multi-channel filtering like most single channel noise reduction methods. To obtain accurate estimation of noise statistics, a number of methods have been proposed for single channel algorithms. Recently, several statistical model based sequential noise estimation methods have been successfully applied to the non-stationary noise estimation for robust speech recognition [3–5]. The statistical-based methods are usually processed in the (log) spectral domain or cepstral domain due to the modeling of clean speech signal (Gaussian mixture model). Hence, they are suitable for the feature compensation in general speech recognition systems. However, it is difficult and computationally inefficient to apply these methods to the QR decomposition based multi-channel filtering, because the multi-channel filtering is processed in the time domain and its main purpose is to estimate the enhanced speech signal waveform. In

* Corresponding author. Tel.: +82 2 880 1774; fax: +82 2 883 2210.
E-mail address: kgb@ispl.snu.ac.kr (G. Kim).
[1] The authors are also affiliated with the Institute of New Media and Communications (INMC).

0003-682X/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apacoust.2007.03.005

practice, it would be better to recursively estimate the noise by controlling the forgetting factor in a conventional multi-channel filter scheme.

In conventional multi-channel filtering, the noise statistics are recursively estimated with a forgetting factor during noise-only periods and kept unchanged during speech-present periods by a voice activity detector and a fixed forgetting factor. Instead of using an explicit VAD for noise estimation, we propose a time varying forgetting factor parameterized as a function of NCC between microphones, which has been used for finding the direction of signals [6]. If we assume that the direction of speech source is known and the time delays between microphones are compensated, the spatial coherence measure such as the NCC is a reliable measure for the speech presence probability even in low SNRs as long as the speech and the noise sources are physically located at different positions. The NCCs of two different microphones are estimated from the signals in a given temporal window, and are averaged over all microphone pairs for the robustness against the mismatch in microphones and array perturbations. The averaged NCC is smoothed by the first-order recursive averaging, and is mapped to a smoothing factor that determines the degree of smoothing. The objective evaluation confirms that the proposed method improves the performance of noise reduction.

2. Multi-channel optimal filtering

The signal model for multi-channel filtering is described as

$$x_i(k) = d_i(k) + v_i(k) \;\; i = 1, \ldots, M \tag{1}$$

where $x_i(k)$ denotes the observed signal at the ith microphone at time k, $d_i(k)$ is the desired speech signal, $v_i(k)$ is the additive noise component, and M is the number of microphones. The stacked data vector $\mathbf{x}(k)$ is defined as

$$\mathbf{x}(k) = \begin{pmatrix} \mathbf{x}_1(k) \\ \mathbf{x}_2(k) \\ \vdots \\ \mathbf{x}_M(k) \end{pmatrix}, \quad \mathbf{x}_i(k) = \begin{pmatrix} x_i(k) \\ x_i(k-1) \\ \vdots \\ x_i(k-N+1) \end{pmatrix} \tag{2}$$

where N is the length of filter tap for each channel. The noise vector $\mathbf{v}(k)$ and the desired speech signal vector $\mathbf{d}(k)$ are defined in the same manner.

We assume that the speech and noise signals are uncorrelated, which makes it possible to estimate the speech correlation matrix from the observed signal and noise. In addition, we also assume that the noise signal is rather stationary as compared to the speech signal so that the noise correlation matrix can be estimated during noise-only periods. With these assumptions, the MMSE optimal filter for multi-channel noise reduction can be written as

$$\begin{aligned} \mathbf{W}_{\mathrm{wf}}(k) &= (E\{\mathbf{x}(k)\mathbf{x}^{\mathrm{T}}(k)\})^{-1} E\{\mathbf{x}(k)\mathbf{d}^{\mathrm{T}}(k)\} \\ &\simeq (E\{\mathbf{x}(k)\mathbf{x}^{\mathrm{T}}(k)\})^{-1} (E\{\mathbf{x}(k)\mathbf{x}^{\mathrm{T}}(k)\} - E\{\mathbf{v}(k)\mathbf{v}^{\mathrm{T}}(k)\}). \end{aligned} \tag{3}$$

In Eq. (3), we can estimate the correlation matrix using the sample correlation matrix as

$$E\{\mathbf{x}(k)\mathbf{x}^{\mathrm{T}}(k)\} \simeq \mathbf{X}^{\mathrm{T}}(k)\mathbf{X}(k) \tag{4}$$

where $\mathbf{X}(k)$ is the input data matrix. If we assume that the sample correlation matrix is recursively updated with an appropriate exponential weighting factor λ_x as

$$\mathbf{X}^{\mathrm{T}}(k+1)\mathbf{X}(k+1) = \lambda_x^2 \mathbf{X}^{\mathrm{T}}(k)\mathbf{X}(k) + (1-\lambda_x^2)\mathbf{x}(k+1)\mathbf{x}^{\mathrm{T}}(k+1), \tag{5}$$

then the input data matrix is defined as

$$\mathbf{X}(k+1) = \begin{bmatrix} \sqrt{1-\lambda_x^2}\,\mathbf{x}^{\mathrm{T}}(k+1) \\ \lambda_x \mathbf{X}(k) \end{bmatrix}. \tag{6}$$

We can also define the noise data matrix $\mathbf{V}(k)$ and the sample correlation matrix for noise $\mathbf{V}^{\mathrm{T}}(k)\mathbf{V}(k)$ in the same manner from the noise vector $\mathbf{v}(k)$. The QR decomposition of $\mathbf{X}(k)$ leads to a recursive implementation of the multi-channel optimal filter [2].

3. Time varying forgetting factor for the noise estimation

Instead of a conventional noise estimation based on hard boundary VAD, we propose an adaptive estimation method using the time varying forgetting factor as

$$\begin{aligned} \mathbf{V}^{\mathrm{T}}(k+1)\mathbf{V}(k+1) &= \lambda_v^2(k+1)\mathbf{V}^{\mathrm{T}}(k)\mathbf{V}(k) \\ &\quad + (1-\lambda_v^2(k+1))\mathbf{x}(k+1)\mathbf{x}^{\mathrm{T}}(k+1). \end{aligned} \tag{7}$$

During noise-only periods, the forgetting factor $\lambda_v(k)$ is usually close to 1 so that the noise reduction performance depends mainly on the spatial characteristics of the input signals and long term averaged spectral characteristics. In this way we can also suppress spectrally non-stationary noise without short-time effects such as musical noise. $\lambda_v(k)$ can be parameterized by the NCC, provided that the direction of the speech source is known and the time delays between microphones are compensated, i.e., the array is steered to the direction of the speech source. The NCC at time k between the mth and the nth microphone is defined as

$$\rho_{mn}(k) = \frac{E\{x_m(k)x_n(k)\}}{\sqrt{E\{x_m^2(k)\}E\{x_n^2(k)\}}}, \tag{8}$$

which tends to be close to 1 for the signals in the look direction and has lower values for the signals from other directions. We can estimate the NCC between two microphones for a given time window by taking the inner product of two normalized signal vectors. The NCCs between all microphone pairs are also averaged for the robustness against the microphone mismatch and array perturbations. The averaged NCC estimates are smoothed in time by a first-order IIR filtering as

$$\hat{\rho}(k) = \frac{2}{M(M-1)} \sum_{m=1}^{M-1} \sum_{n=m+1}^{M} \frac{x_m^{\mathrm{T}}(k)x_n(k)}{\|x_m(k)\| \cdot \|x_n(k)\|} \tag{9}$$

$$\tilde{\rho}(k+1) = \lambda_\rho \tilde{\rho}(k) + (1-\lambda_\rho)\hat{\rho}(k) \tag{10}$$