Digital Signal Processing 25 (2014) 123–133

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Hybrid probabilistic adaptation mode controller for generalized sidelobe cancellers applied to multi-microphone speech enhancement

Seon Man Kim [a], Hong Kook Kim [b,*]

[a] Institute of Sound and Vibration Research, University of Southampton, Southampton SO17 1BJ, UK
[b] School of Information and Communications, Gwangju Institute of Science and Technology, 1 Oryong-dong, Buk-gu, Gwangju 500-712, Republic of Korea

ARTICLE INFO

Article history:
Available online 15 November 2013

Keywords:
Multiple-microphone speech enhancement
Generalized sidelobe canceller
Adaptation mode controller
Adaptive beamformer
Speech presence probability

ABSTRACT

In this paper, we propose a new adaptation mode controller (AMC) for a generalized sidelobe canceller (GSC)-based speech enhancement system. Here, a likelihood ratio for target speech presence was first estimated and then utilized to estimate both the local target speech presence probability (SPP) and global SPP. Next, the estimated SPPs were applied to the design of an AMC that controlled the parameters of adaptive filters for an adaptive blocking matrix (ABM) and noise canceller (NC). In particular, the combination of local and global SPPs was applied to the AMC in the ABM, whereas only global SPPs were used for the NC. Finally, a multiple-microphone speech enhancement system was constructed on the basis of a GSC having the proposed AMC. The performance of the speech enhancement system was subsequently evaluated in terms of the perceptual evaluation of speech quality (PESQ) and the cepstral distortion (CD) for car noise conditions. It was shown from this evaluation that a speech enhancement system using the proposed AMC method provided better performance than conventional AMC methods using power ratios between the target and non-target directional signals, the inter-channel normalized cross-correlation, and the local SPPs only.

1. Introduction

The need for robust performance in adverse noise environments has led to advances in speech enhancement research [1]. Recently, techniques for multiple-microphone systems have been reported, as opposed to techniques for single-microphone systems, which have also demonstrated improved performance in terms of speech quality [1,2]. Among them, beamforming has become attractive because it enables the extraction of the desired target directional signals that are contaminated by non-target directional noise [2].

A generalized sidelobe canceller (GSC) is one of the most popular beamformers because it has structural simplicity and is easy to implement [2,3]. A GSC mainly consists of a fixed beamformer (FBF), blocking matrix (BM), and noise canceller (NC). The FBF provides a fixed sound beam in the target direction so that non-target directional signals can be attenuated. In contrast, the BM blocks the target directional signal so that only non-target directional signals can pass through. The NC then uses adaptive filtering to enhance the target directional signal obtained from the FBF and BM.

Among the three components of a GSC, the BM plays a main role. To be specific, GSCs suffer from unsuitable phase differences between the multiple-microphone signals for the target direction [2–5]. Such unsuitable phase differences or phase errors are unavoidable because the theoretical sound propagation model in the GSC does not always reflect the real environment. These problems have several potential sources such as the microphone position, the microphone characteristics, and the mismatch between the true target direction-of-arrival (DOA) and the assumed DOA. Since phase errors lead to target directional speech leakage in the BM output signal, it results in target speech cancellation at the GSC output [4,5]. In order to mitigate this problem, an adaptive filter is typically employed in the BM, which is referred to as an adaptive BM (ABM) [4,5].

Even though an adaptive filter in the ABM is designed to alleviate the influence of phase errors, its performance is not guaranteed in a real environment, as incorrect adaptation of the ABM may result in incomplete blocking of the target directional signal. To overcome this problem, the use of an adaptation mode controller (AMC) has been proposed, in which filter adaptation is alternately carried out depending on the target speech activation [6]. To control filter adaptation, the AMC determines the target speech activation in different ways such as a hard decision or a soft decision. A hard-decision AMC classifies target speech activation into one of two binary hypothesis models, i.e., target speech absence or presence [6]. In contrast, a soft-decision AMC provides a probabilistic value ranging from zero to one for the presence of target speech; in [7], it was reported that a probabilistic soft-decision AMC performed better than a hard-decision AMC.

* Corresponding author. Fax: +82 62 715 2204.
E-mail address: hongkook@gist.ac.kr (H.K. Kim).

1051-2004/$ – see front matter
http://dx.doi.org/10.1016/j.dsp.2013.11.005

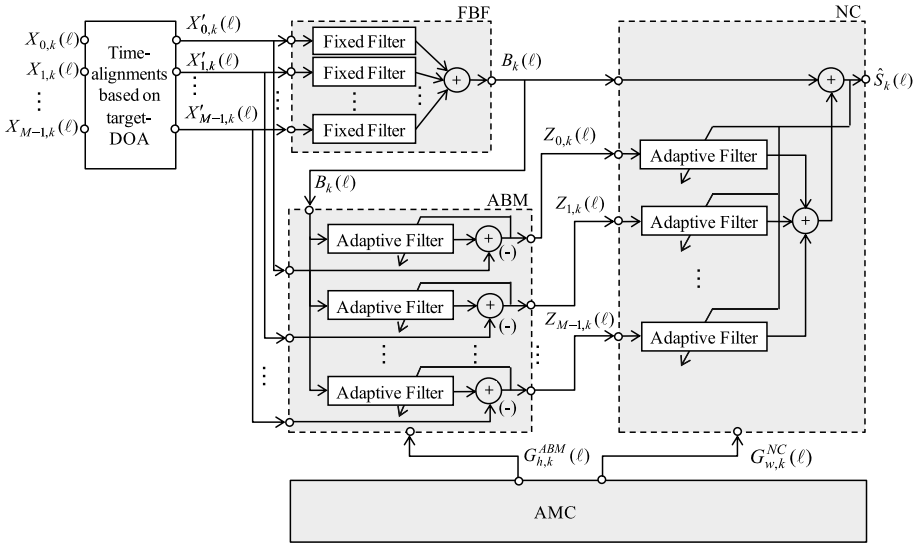

Fig. 1. Block diagram of a GSC-based speech enhancement system with an AMC.

To select the binary or soft value for the AMC, conventional AMC methods utilize parameters such as the target-to-non-target directional power ratio [6], correlation [8], inter-channel normalized cross-correlation (INCC) [9], and target speech presence probability (SPP) [7]. Among them, the AMC based on the target SPP estimate is often preferred because it can provide a value ranging from zero to one. This implies that reliable estimation of the target SPP becomes crucial for the AMC. The target SPP can be estimated locally in each frequency bin or globally across all frequency bins, referred to as a local decision or a global decision, respectively. It is noted here that conventional probabilistic soft-decision AMCs are based on the local SPP estimate [7], and though this estimate can provide more reliable information about the target speech activation in each frequency bin, the global SPP estimate is primarily useful for detecting noise intervals [10,11]. Therefore, it is expected that the hybrid use of both SPP estimates can provide improved AMC performance compared to the use of either the global or local SPP estimate alone.

In this paper, we propose a new AMC based on the hybrid use of local and global SPP estimates. In our previous work, it was shown that the hybrid use of local and global SPP estimates to update the adaptive filter coefficients in the ABM and NC improved the AMC performance [10]. In particular, the local and global SPP estimates were simply multiplied for the AMC. However, it was difficult to confirm whether the influences of the two SPP estimates were optimized for AMC performance because such a combination gave equal emphasis to the global and local SPP estimates for AMC performance. Instead, a more general form for combining two SPP estimates is proposed here in order to further improve the AMC performance; i.e., the proposed method can provide different weightings to each SPP estimate to better control the AMC.

The remainder of this paper is organized as follows. Following this introduction, conventional AMCs for a GSC-based speech enhancement system are briefly reviewed in Section 2. Section 3 describes the overall procedure of the GSC-based speech enhancement system with the proposed AMC. After that, a general form for combining the global and local SPP estimates is proposed. Section 4 then evaluates the performance of a GSC-based speech enhancement system employing the proposed AMC in terms of the perceptual evaluation of the speech quality (PESQ) and the cepstral distortion (CD) in car noise environments. Finally, our findings are summarized in Section 5.

2. Conventional AMCs for a GSC-based speech enhancement system

Fig. 1 shows a block diagram of a frequency-domain GSC with an AMC. For given M-channel multiple-microphone signals, $x_m(n)$ for $m = 0, 1, \ldots, M-1$, we apply a short-time Fourier transform (STFT) in order to obtain its spectral representation, $X_{m,k}(\ell)$, where k $(= 0, 1, \ldots, K-1)$ is a frequency bin and ℓ is a frame index. Note that K denotes the total number of frequency bins. In the figure, the FBF output signal, $B_k(\ell)$, is first obtained by filtering and summing all of the time-aligned versions $X'_{m,k}(\ell)$ of $X_{m,k}(\ell)$. Simultaneously, the ABM transforms $X'_{m,k}(\ell)$ into target directional blocked signals, $Z_{m,k}(\ell)$. Next, $B_k(\ell)$ and $Z_{m,k}(\ell)$ are converted into a target directional enhanced signal, $\hat{S}_k(\ell)$, using an adaptive filter in the NC.

Let $S_k(\ell)$ and $N_k(\ell)$ be the k-th spectral components at the ℓ-th frame of the target speech source and additive noise, respectively, where $S_k(\ell)$ is assumed to be uncorrelated with $N_k(\ell)$ [1,11–13]. Then, $X_{m,k}(\ell)$ is represented as

$$X_{m,k}(\ell) = S_{m,k}(\ell) + N_{m,k}(\ell) \tag{1}$$

where $S_{m,k}(\ell)$ and $N_{m,k}(\ell)$ are the target speech and noise recorded at the m-th microphone, respectively, and they are uncorrelated each other. In (1), $S_{m,k}(\ell)$ is a function of $S_k(\ell)$ and it is represented by taking into account the propagation delay and propagation gain according to the actual microphone array geometry such that $S_{m,k}(\ell) = S_k(\ell) \cdot a_{m,k} \exp(-j\omega_k \bar{\tau}_m)$, where $a_{m,k}$ is the propagation gain of $S_k(\ell)$ at the m-th microphone and $\bar{\tau}_m$ is the time delay of $S_k(\ell)$. In addition, ω_k is the angular frequency in radians at the k-th frequency bin. Here, $a_{m,k}$ becomes 1 by assuming that the target source is located far enough from the microphone array, referred to as the acoustic far-field condition [2]. Moreover, the propagation time delay at a reference microphone (the 0-th microphone in this study) is assumed to equal zero, i.e., $\bar{\tau}_0 = 0$, leading the relationship of $S_{0,k}(\ell) = S_k(\ell)$. Thus, $S_{m,k}(\ell)$ is also represented as

$$S_{m,k}(\ell) = S_k(\ell) \cdot \exp(-j\omega_k \tau_m) \tag{2}$$

where τ_m is the time difference of arrival (TDOA) of $S_k(\ell)$ between the m-th microphone and the reference one that is the 0-th microphone in this paper. Note here that for a given microphone configuration, τ_m can be estimated using a localization algorithm such as the generalized cross-correlation (GCC) or steered response