Speech Enhancement Using Improved Generalized Sidelobe Canceller in Frequency Domain with Multi-channel Postfiltering

Kai Li, Qiang Fu and Yonghong Yan

ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, No.21 Beisihuan Xilu, Beijing, 100190, China
{likai,qfu,yyan}@hccl.ioa.ac.cn

Abstract
In this paper, we propose a speech enhancement algorithm which has the feature of interaction between adaptive beamforming and multi-channel postfilter. A novel subband feedback controller based on speech presence probability is applied to Generalized Sidelobe Canceller algorithm to obtain a more robust adaptive beamforming in adverse environment and alleviate the problem of signal cancellation. A multi-channel postfiltering is used not only to further suppress diffuse noises and some transient interferences, but also to give the speech presence probability information in each subband. Experimental results show that the proposed algorithm achieves considerable improvement on signal preservation of the desired speech in adverse noise environments, consisting of both directional and diffused noises over the comparative algorithms.

Index Terms: Speech Enhancement, Microphone Array, Generalized Sidelobe Canceller, Subband Feedback Control, Postfilter

1. Introduction
Microphone array has been widely used to improve the performance of speech communication and automatic speech recognition (ASR) systems because of their effectiveness in enhancing the quality of the captured speech. Compared with single channel systems, a substantial gain in performance is obtainable due to the spatial filtering capability to suppress interfering signals coming from undesired directions.

To suppress directional noises, a lot of algorithms based on beamformer have been proposed [1]. The Frost beamformer [2] was one of the first array structures to handle adaptive broadband array processing. Griffiths and Jim [3] proposed an alternative method of Frost's algorithm and introduced the generalized sidelobe canceller (GSC) solution, which not only effectively reduces the computational complexity but also provides flexibility to implement different beamformers. However, GSC algorithm suffers from signal cancellation problem because of the steering vector error, reverberation or imperfect microphones[1, 4]. Besides, the adaptive filter in GSC is not robust in transient noise environment. In order to prevent the algorithms from diverging, several trials need to be conducted before a proper step-size is found. These drawbacks obviously will obstruct the use of these adaptive beamforming algorithms in practice.

To suppress diffuse noises, postfilter is normally needed. Zelinski’s postfilter [5] employs auto- and cross-correlation functions of received multi-channel signals to derive a proper gain for enhancement. However, this method is based on the assumption of incoherent noise field which is seldom satisfied in practical environments. A generalized expression for Zelinski postfilter has been derived based on the a priori knowledge of noise field [6]. Recently, Israel Cohen introduced a multichannel signal presence probability estimation based postfilter which is particularly advantageous in nonstationary noise environments[7].

In this paper, an improved Generalized Sidelobe Canceller algorithm which has the features of interaction between beamforming and multi-channel postfilter is proposed, as shown in Figure1. The outputs of Fixed Beamforming (FBF) and a modified Blocking Matrix (BM) which uses more spatial information are analyzed in the STFT domain and regrouped into auditory subbands according to the Bark scale, which mimics the auditory characteristics of human ears. And adaptive interference cancellation is performed in each subband.

A closed-loop controller uses feedback to control states of a dynamic system can keep the control error to a minimum and dynamically compensate for disturbances to the system[8]. In speech enhancement area, adaptive beamforming can be seen as a dynamic system which is adaptive to the adverse environment. Based on this consideration, we propose a novel subband feedback controller based on speech presence probability which is derived from the postfilter to feedback control the adaptive interference canceller of GSC in each subband. We modified Cohen’s multi-channel postfilter so that signal presence probability (SPP) in each auditory subband can be derived. The update of the filter coefficients is slowed down when the desired speech is present so that the proposed algorithm is more robust to array imperfection or reverberation, as the desired speech may leak into the reference channel. The interaction between the multichannel processing and the postfilter leads to better signal preservation thus improves the algorithm’s overall performance.

2. Proposed Speech Enhancement Algorithm
Consider a four-sensor microphone array in noisy environment, the observed signal on each microphone is composed of desired speech signal, directional noises arriving from determinable directions and diffuse noises propagating in all directions. The aim of our task is to reduce both directional and diffuse noises simultaneously while keeping the desired speech distortionless. To implement this idea, we construct a noise reduction system, shown in Fig.1, which consists of three main parts: directional noise suppression, diffuse noise suppression, and the interaction of these two parts through a subband feedback controller, detailed in the following.
2.1. Directional Noises Suppression

To suppress directional noises, we proposed an improved Generalized Sidelobe Canceller algorithm. In the original GSC beamformers, the BM parts were implemented by subtracting between observed signals on adjacent sensors, which indicates that only limited spatial information was used. Comparatively, the modified BM considers the spatial information not only between adjacent sensors but also other sensor pairs, given by:

\[
\begin{bmatrix}
1 & -1 & 0 & 0 \\
1 & 0 & -1 & 0 \\
1 & 0 & 0 & -1
\end{bmatrix}
\]

Experiment demonstrates the effectiveness of this BM in [3].

Signal from the output of fixed beamformer (FBF), and Block Matrix (denoted as \(y(n)\) and \(u_m(n)\) respectively) are segmented into temporal frames and analyzed by short-time Fourier transform (STFT).

\[
y(n) \xrightarrow{\text{STFT}} Y(k, l), \quad u_m(n) \xrightarrow{\text{STFT}} U_m(k, l) \quad (1)
\]

in which \(l\) and \(k\) denote the index of temporal frames and frequency bins, respectively, \(m = 1, 2, \ldots, M - 1, M\) is the number of the microphones.

We regroup the frequency domain signal of each frame into \(B\) groups according to Bark scale. The vectors of bins within the \(b\)th group are denoted as \(Y^b(l)\) and \(U^b_m(l)\) respectively.

Recall that our goal is to minimize the output power under a constraint on the response at the desired direction. Since the constraint is satisfied in the fixed beamformer, this is an unconstrained minimization similar to Widrow’s classical Adaptive Noise Cancellation problem [9].

\[
J^b_m(l) = E[\|Y^b(l) - W^b_m(l)U^b_m(l)\|^2] \quad (2)
\]

\(J^b_m(l)\) denotes the energy in \(b\)th band, \((\cdot)^H\) is the Hermitian transpose operator and \(E[\cdot]\) is the expectation operator. Minimizing \(J^b_m(l)\) leads to

\[
W^b_m(l)_{\text{opt}} = \frac{\Phi^b_{U_mY}(l)}{\Phi^b_{U_mU_m}(l)} \quad \text{if} \quad \frac{d\Phi^b_{U_mU_m}(l)}{dW^b_m(l)} = 0 \quad (3)
\]

where

\[
\Phi^b_{U_mY}(l) = E[U^b_m(l)(Y^b(l))^H] \quad (4)
\]

\[
\Phi^b_{U_mU_m}(l) = E[U^b_m(l)(U^b_m(l))^H] \quad (5)
\]

In order to track changes, we process the signals by segments. The following unconstrained frequency domain normalized LMS (UFNLMS) algorithm is used. The adaptive interference canceller filter in each of the subband is updated by a modified UFNLMS with a different norm constraint.

\[
W^b_m(l + 1) = W^b_m(l) + \mu \frac{U^b_m(l)(Y^b(l))^H}{P^b_{\text{est}}(l)} \quad (6)
\]

where

\[
P^b_{\text{est}}(l) = \alpha P^b_{\text{est}}(l - 1) + (1 - \alpha) \sum_{m=1}^{M-1} |X^b_m(l)|^2 \quad (7)
\]

For a standard UFNLMS algorithm, we should calculate \(P^b_{\text{est}}(l)\) using the power of the noise reference signals, but we find in experiment that the signal cancellation problem is serious if we update the weight during speech presence, so we use \(X^b_m(l)\) which is the frequency domain representation of input sensor signals instead. The performance is improved due to the fact that the adaptation term becomes relatively small during speech presence. This can be seen as an implicit control of the adaptive filter.

In order to precisely control the filter adaptation in Generalized Sidelobe Canceller, We proposed a method to use the signal presence probability derived from the postfilter to feedback control the adaptive interference canceller of GSC in each subband, which will be detailed in 2.3.

As speech is concerned, the energy of desired signal mainly centralizes in low frequencies, so the signal in this area appears to be more colorful, while in higher frequencies, signal energy appears to be much weaker. So it is reasonable that nonuniform filterbanks, instead of the uniform ones, should be used to make the low frequency bands narrower to proceed explicit analysis.
while in the high frequency bands, the bandwidth should be broader to contain more signal energy in order that the adaptive interference canceller may converge more smoothly. The auditory subband method is proved to be effective in our previous work [10].

2.2. Diffuse Noises Suppression

The residual diffuse noises are further suppressed by a signal presence probability-based multi-channel postfilter [7], which uses a multi-channel soft signal detection based on the nonstationary of the signals and the transient power ratio between the beamformer primary output and its reference noise which to estimate the speech presence probability and noise power spectral density and then an optimal gain function that minimizes the mean square error of the log-spectral amplitude is applied.

The postfilter estimates the EM gain \( G_{EM}(k,l) \) and SPP \( P(k,l) \). And final gain for enhancement \( G(k,l) \) is reached by

\[
G(k,l) = (G_{EM}(k,l))^{P(k,l)}G_{min}^{1-P(k,l)}
\]

where \( G_{min} \) is the minimum gain allowed (typically \(-20\)dB). \( G_{EM}(k,l) \) is derived from single channel approach mainly and is able to reduce the stationary and quasi-stationary noises. And \( P(k,l) \), which suggests the probability of the desired speech exists in the corresponding time-frequency unit, is calculated by considering the ratio between the transient power of the GMC output \( Z(k,l) \) and the transient power of the BM output reference signal \( V_{m}(k,l) \). A low ratio indicates a larger transient power in the reference channel, which means that an interfering source is probably present. In this case, a smaller \( P(k,l) \) is assigned. Thus the non-stationary noise in \( Z(k,l) \) will be further suppressed according to (8) because a small \( P(k,l) \) will make the final gain approach \( G_{min} \).

The system output is given by

\[
S(k,l) = G(k,l) \cdot Z(k,l)
\]

As mentioned in Sect.2.3, SPP in each auditory subband is needed for constraining filter updates. This can be achieved by averaging SPP of the time frequency units within the corresponding subbands.

2.3. Subband Feedback Controlled Adaptive Filters

In practical implementations, the target speaker may not stay precisely at \( \theta \). Moreover, the desired speech will also leak into the reference channel due to echo and reverberation characteristics of the room. Furthermore, the position and frequency response of the microphones may not be as precise as expected, leading to imperfect cancelation of the desired speech in the reference channel. So the minimization of \( J_{in}(l) \) in (2) does not necessarily lead to maximization of output SNR, instead, a certain proportion of speech signal will be canceled as a result. The leakage will also cause false fluctuations of filter coefficients.

To improve the system’s robustness against the adversities mentioned above, it is preferable that the updating rate of the adaptive filters should be controlled according to the presence of the desired speech. When the desired speech is present, update mentioned in (6) should be slowed down.

The adaptation speed and steady state error of the adaptive filter are highly related to the step-size constant, but it is very hard to find the optimal step-size which guarantees the good performance in a general environment. So \( \mu \) in (6) must vary in different bands and temporal frames.

We propose an time-varying step-size which is controlled by the speech presence probability in each subband which is derived from the postfilter described in the next section.

\[
p^b(l) = \frac{1}{M_b} \sum_{i=b_1,b_2,...} P^i(l)
\]

in which \( M_b \) is the number of frequency bins within the \( b \)th subband.

\[
\mu^b(l) = (1 - p^b(l))\mu = (1 - \frac{1}{M_b} \sum_{i=b_1,b_2,...} P^i(l))\mu
\]

\( p^b(l) \) is the signal presence probability derived from the postfilter described in the last section, \( 0 < p^b(l) < 1 \). A greater \( p^b(l) \) indicates a high probability that the desired signal may exist in the \( b \)th subband during the \( l \)th frame. Thus a larger \( \mu^b(l) \) is achieved according to (11), resulting in slow updates of the adaptive filters which preserves the speech components. And a small \( p^b(l) \) means the desired signal is mostly absent. So the updates become fast enough to adapt to the changing nature of the interferences.

3. Evaluations

The microphone array used in this work is composed of 4 omnidirectional MEMS microphones in broadside orientation. The distance between the microphones is set to be 3cm. The system is implemented under a sampling rate of 8kHz.

The experiment took place in a conference room of \( 6m \times 5m \times 3m \) with a reverberation time of 300ms. Two interferences (a competing speaker and a gauss white noise source) are located in 90° and 45° of the array, respectively. The speech source is ten male and ten female TIMIT sentences. The multichannel clean speech is generated by computer simulation in a virtual room [11] with the same size and reverberation time of the conference room in which the interferences are recorded, so that clean speech signal can be obtained for objective measurements. And then we mix the two parts with different global SNR levels(6-6dB). All the sound sources are 1m away from the array.

For comparison, the multichannel noisy speech is processed with six methods listed below.

1. GSC algorithm in time domain (GSC-TD) [3].
2. GSC algorithm in frequency domain (GSC-FD) [12].
3. GSC-FD with modified Blocking Matrix(GSC-FD*).
5. Cohen’s algorithm [13].

The results of the above methods are evaluated in Noise Reduction (NR) and log-spectral distance (LSD) [13], as is shown in Fig.2.

Figure2 shows that adaptive beamforming using frequency-domain adaptive filter exhibits fast convergence behavior and better performance of nulling wideband interferences. We can also notice that the modified fixed BM using more spatial information gains some improvement. The subband feedback controlled method can alleviate the problem of signal cancellation in adaptive beamformer and has a better desired signal preservation. Further to demonstrate this point, PSEQ-MOS is employed, as shown in Table 1.
4. Conclusion

A multi-channel speech enhancement algorithm is proposed. The algorithm consists of three parts: directional noise suppression, which is based on an improved Generalized Sidelobe Canceller with subband feedback controlled adaptive filters; diffuse noise suppression which is implemented by a multi-channel postfiltering based on speech presence probability; and the interaction of adaptive beamforming and postfiltering through a subband feedback controller. Experimental results indicate that the subband feedback controller make the filter adaptation more robust and alleviate the problem of signal cancellation in adaptive beamformer. The proposed algorithm achieves considerable improvement on signal preservation of the desired speech in adverse noise environments over the comparative algorithms. Considering that the speech presence probability used by subband feedback controller is obtained from the postfilter, it does not increase many computational cost.

5. Acknowledgement

This work is partially supported by The National Science&Technology Pillar Program (2008BA150B03), National Natural Science Foundation of China (NO.10574140, 10925419, 90920302,10874203, 60875014).

6. References


