Robust Relative Transfer Function Estimation for Dual Microphone-Based Generalized Sidelobe Canceller

Kihyeon KIM†, Nonmember and Hanseok KO†a), Member

SUMMARY In this Letter, a robust system identification method is proposed for the generalized sidelobe canceller using dual microphones. The conventional transfer-function generalized sidelobe canceller employs the non-stationarity characteristics of the speech signal to estimate the relative transfer function and thus is difficult to apply when the noise is also non-stationary. Under the assumption of W-disjoint orthogonality between the speech and the non-stationary noise, the proposed algorithm finds the speech-dominant time-frequency bins of the input signal by inspecting the system output and the inter-microphone time delay. Only these bins are used to estimate the relative transfer function, so reliable estimates can be obtained under non-stationary noise conditions. The experimental results show that the proposed algorithm significantly improves the performance of the transfer-function generalized sidelobe canceller, while only sustaining a modest estimation error in adverse non-stationary noise environments.

key words: dual microphone arrays, generalized sidelobe canceller, relative transfer function estimation, spectral classification, W-disjoint orthogonality

1. Introduction

The generalized sidelobe canceller (GSC) is one of the most efficient beamforming algorithms employed to enhance the target speech signal under various noise conditions, such as reverberant office spaces or automotive enclosures [1]–[3]. The process is composed of three parts: a fixed beamformer (FBF) that satisfies the desired constraint on the target, a blocking matrix (BM) that produces the reference noise signal by subtracting pairs of time-aligned input signals, and an adaptive noise canceller (ANC) that cancels the noise in the FBF output [4].

The original GSC, however, has an inherent limitation in that it assumes delay-only signal propagation. Gannot et al. addressed this problem and derived a GSC solution based on the relative transfer functions (RTFs) between sensors in response to a desired speech signal [5], [6]. The transfer-function GSC (TFGSC) algorithm carefully specifies the FBF and BM by using the RTFs and shows improved performance in enhancing the speech signal. However, since the RTF estimation employs the non-stationary characteristics of the speech signal under the assumption that the noise is stationary [7], the TFGSC cannot be directly applied in non-stationary noise environments.

In this Letter, a robust RTF estimation algorithm is proposed for the dual-channel TFGSC in the presence of non-stationary noise. Basically, we focus on the non-stationary noise whose signal distributions are sparse in the time-frequency domain, such as an interfering human voice, and assume W-disjoint orthogonality (WDO) between the speech and the non-stationary noise. The proposed algorithm introduces a target dominance indicator (TDI) function to discriminate the time-frequency bins, where the desired speech signal dominates the interference, and then performs adaptation to obtain a reliable RTF estimate with these target-dominant input components.

2. Conventional RTF Estimation for TFGSC

Let \(s(t)\), \(a_i(t)\) be the desired speech signal and the impulse response of the path from the speech source to the \(i\)-th microphone. When the \(i\)-th sensor includes the interfering noise, \(n_i(t)\), the observed signal \(z_i(t)\) is given by

\[
z_i(t) = a_i(t) \ast s(t) + n_i(t), \quad i = 1, 2
\]

where \(\ast\) denotes convolution. Under the assumption that the room impulse responses are time-invariant and the analysis frame duration is chosen for the signal to be considered stationary over the analysis frame, the sensor signal can be represented in the short time Fourier transform (STFT) domain as

\[
Z_i(k, l) = A_i(k)S(k, l) + N_i(k, l)
\]

where \(A_i\) denotes the general transfer function of the impulse response and \((k, l)\) the frequency-frame index. In dual-channel TFGSC, the FBF, BM, and ANC blocks generate the primary signal \((Y)\), the reference noise signal \((V)\), and the output signal \((\hat{S})\), respectively, as shown in the following equations [6]:

\[
Y(k, l) = [1 + |H(k)|^2]^{-1} [Z_1(k, l) + H^*(k)Z_2(k, l)],
\]

\[
\hat{Y}(k, l) = Z_2(k, l) - H(k)Z_1(k, l),
\]

\[
\hat{S}(k, l) = Y(k, l) - G^*(k, l)V(k, l),
\]

where \(H\) is the RTF defined by

\[
H(k) = A_2(k)/A_1(k),
\]

and \(G\) is the ANC filter adjusted to minimize the output power.

For the TFGSC, the conventional RTF estimation is based on the non-stationarity of the desired signal. Let \(\Phi_{Z_1Z_1}\) be the cross-power spectral density (CPSD) between \(Z_1\) and
Then, the CPSD can be expressed as
\[
\Phi_{Z,Z}(k, l) = H(k)\Phi_{Z,Z}(k, l) + \Phi_{VZ}(k, l)
\] (7)
from (4). If the noise signal \(N_i\) \((i = 1, 2)\) is assumed to be stationary and the speech signal \(S\) is independent of \(N_i\), then \(\Phi_{VZ}(k, l)\) is independent of \(l\): \(\Phi_{VZ}(k, l) = \Phi_{VZ}(k)\). However, \(\Phi_{VZ}(k, l)\) actually has the estimation error of \(\varepsilon(k, l) = \hat{\Phi}_{VZ}(k, l) - \Phi_{VZ}(k)\) when the reference noise signal \(V\) includes the speech leakage. This error changes over time due to the non-stationarity characteristics of the speech signal. Therefore, an unbiased estimate for \(H\) is obtained by solving the least square (LS) problem of the following over-determined equation sets [6], [7]:
\[
\begin{bmatrix}
    \hat{\Phi}_{Z,Z}(k, 1) \\
    \hat{\Phi}_{Z,Z}(k, 2) \\
    \vdots \\
    \hat{\Phi}_{Z,Z}(k, L)
\end{bmatrix} = \begin{bmatrix}
    \hat{\Phi}_{Z,Z}(k, 1) & 1 \\
    \hat{\Phi}_{Z,Z}(k, 2) & 1 \\
    \vdots & \vdots \\
    \hat{\Phi}_{Z,Z}(k, L) & 1
\end{bmatrix} \begin{bmatrix}
    H(k) \\
    \Phi_{VZ}(k)
\end{bmatrix} + \begin{bmatrix}
    \varepsilon(k, 1) \\
    \varepsilon(k, 2) \\
    \vdots \\
    \varepsilon(k, L)
\end{bmatrix}
\] (8)
where \(\hat{\Phi}_{Z,Z}\) represents the corresponding CPSD estimate and \(L\) is the number of frames within the analysis interval.

This approach provides reliable RTF estimates in stationary noise environments. However, in many cases, noise signals are highly non-stationary, such as interfering human voices or TV sounds and, thus, the LS based estimation method is not available any more.

3. Proposed Algorithm

To obtain a robust RTF estimate under non-stationary noise conditions, the proposed algorithm assumes the WDO between the speech signal and the non-stationary noise. If the non-stationary noise signal has a sparse time-frequency representation, such as in the case of an interfering human voice, the WDO condition can be applied to the target speech and the noise, which supports the mutually disjoint expression in the time-frequency domain [8], as follows:
\[
S(k, l)N(k, l) \approx 0, \forall k, l.
\] (9)
Equation (9) states that the energy of the noise is approximately zero when the speech signal is dominant at a time-frequency bin. This condition makes the estimation error \(\varepsilon(k, l) = \hat{\Phi}_{VZ}(k, l) - \Phi_{VZ}(k, l)\) be the CPSD estimate between the speech leakage and the first sensor’s speech component. That is, the estimation problem can be simplified into the minimization of \(\Phi_{VZ}\) over the speech-dominant time-frequency bins (SDTFBs). Hence, we can establish the optimization criterion based on the minimum mean square algorithm:
\[
\hat{H}(k) = \arg \min_{\hat{H}(k)} E \left[ |I(k, l)\hat{\Phi}_{VZ}(k, l)|^2 \right]
\] (10)
where \(E[\cdot]\) denotes the mathematical expectation and \(I\) denotes the TDI function used to discriminate SDTFBs. A recursive online solution based on the normalized least mean square (NLMS) algorithm is given by
\[
\hat{H}(k, l + 1) = \hat{H}(k, l) + \mu\hat{\Phi}_{Z,Z}^\dagger(k, l)I(k, l)e(k, l)
\] (11)
where \(\mu\) is the learning rate and \(e\) is the error signal:
\[
e(k, l) = \hat{\Phi}_{Z,Z}(k, l) - \hat{H}(k, l)\Phi_{VZ}(k, l).
\] (12)
Then, the TDI function \(I(k, l)\) is effectively designed with the inter-microphone time delay \(\delta\) and the TFGSC output \(\hat{S}\) in the following way:
\[
I(k, l) = \begin{cases} 
  1 & \text{if } |\delta(k, l)| < \tau \text{ and } |\hat{S}(k, l)| > C_{SNR} \cdot \eta(k) \\
  0 & \text{otherwise}
\end{cases}
\] (13)
where \(\tau\) is the pre-defined threshold and the time delay, \(\delta\), is given in the same way as [9]:
\[
\delta(k, l) = -\frac{K}{2nk} \frac{Z_2(k, l)}{\hat{H}(k, l)Z_2(k, l)}.
\] (14)
where \(K\) denotes the FFT size. If the desired source is located in a look direction satisfying the RTF of \(H\), the constraint on the time delay in Eq. (13) implies that the desired signal has a bounded incoming angle with respect to the desired direction. The parameter \(\eta\) is given by time averaging the GSC’s output magnitude spectra for the non-speech duration. By comparing the magnitude of the enhanced speech signal with \(\eta\), we can obtain meaningful information as to whether the input signal is target-dominant or not at a given time-frequency bin. The constant \(C_{SNR}\) is employed to control the reliability of the TDI function according to the level of the non-stationary noise. Most of all, a false alarm should be constrained to keep the error variance small in the RTF estimation. That is, the RTF is robustly estimated by preventing inclusion of non-speech time-frequency bins in Eq. (10). However, this constraint can also cause low detection rate of SDTFBs, so makes the RTF estimation to be executed with a smaller number of real SDTFBs. Table 1 shows the \(C_{SNR}\) values and their corresponding detection probabilities when the false alarm is constrained to 0.1 at every signal-to-ratio (SNR) level. Then, detection probabilities are obtained by evaluating hit rate of the proposed TDI among real SDTFBs.

| Input SNR | -5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
|-----------|-------|------|------|-------|-------|-------|
| \(C_{SNR}\) | 1.2   | 1.3  | 1.4  | 1.5   | 1.7   | 2.0   |

| Detection Prob | 0.56 | 0.65 | 0.74 | 0.81  | 0.86  | 0.88  |

Table 1 \(C_{SNR}\) values and detection probabilities at various SNR levels (\(\tau = 0.5\)).
In Table 1, the TDI provides a detection probability of up to 0.88, while maintaining a value of 0.56 at −5 dB. In adverse noise conditions, the TDI has relatively low detection probabilities and the data sparse problem occurs in the RTF estimation. However this problem can be solved by slightly increasing the amount of speech data. For example, if the RTF estimate is converged within 1 second at input SNR of 20 dB, it only requires about 1.6 (≈ 0.88/0.56) seconds at −5 dB.

Finally, in Fig. 1, the overall RTF estimation scheme integrated with the TFGSC is explicitly described.

4. Experiments and Results

To evaluate the performance of the proposed algorithm, we used the signals recorded at a rate of 8 kHz and 16 bits per sample in a typical office room with dimensions of 4 m × 3.5 m × 2.5 m. The reverberation time of the room is about 280 ms. Dual microphones were located with spacing of 10 cm in the middle of the enclosure. The desired speaker was located at a distance of 50 cm in the forward direction (90°) from the microphone array. The non-stationary noise source was located at 70 cm from the array center, along the 45° line. The desired speech signals were 10 Korean sentences of about 3−4 s each and the non-stationary noise signal was of an arbitrary speaker in a lecture setting. These signals were recorded separately and were mixed to generate input signals at various SNR levels ranging from −5 to +20 dB. Time-frequency analysis was performed with a hamming window of 64 ms in length and a 512-point FFT was used for every 32 ms. The PSD estimates at a time-frequency bin are obtained by applying a first order smoothing to the periodogram of the input signals. The performance of the proposed algorithm was verified by two objective measures: the normalized square deviation (NSD) of the RTF in Decibel level defined as

\[ NSD_{dB} = 10 \log_{10} \frac{\sum_{k} |H(k) - \hat{H}(k)|^2}{\sum_{k} |H(k)|^2} \]  

and the perceptual evaluation of speech quality (PESQ) score in [10]. The optimal RTF \( H \) is given by the standard system identification method which uses only clean speech inputs [6]. In Table 2, the proposed algorithm shows a small NSD which is below −9.8 dB in the overall SNR range, while the conventional algorithm produces very poor estimates under adverse non-stationary noise conditions. Table 2 also demonstrates that the proposed algorithm always has a better estimation performance than the NBRE if the given TDI information is ideal.

Tables 3 and 4 show that by using the improved estimates of Table 2 the TFGSC enhance the speech quality in terms of output SNR and PESQ, respectively. In Table 3, the output SNR of the TFGSC typically depends on the performance of the RTF estimation in Table 2. Moreover, the proposed method tends to show further improved results in
Let us now inspect the time delay between the dual microphone inputs and the GSC output. The RTF is estimated by using only the target-dominant time-frequency bins, so that the estimate is robust to the non-stationary noise. The experimental results confirm that the proposed algorithm produces reliable RTF estimates and enhances the performance of the TFGSC in non-stationary noise environments.

### Acknowledgments

This research was supported by Grant R01-2006-000-11162-0 (2008) from the BRPKS and EF of the MSC.

### References

1. K.M. Buckley, “Broad-band beamforming and the generalized sidelobe canceller,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-34, no.5, pp.1322–1323, Oct. 1986.
2. D.V. Compernolle, W. Ma, F. Xie, and M.V. Diest, “Speech recognition in noisy environments with the aid of microphone arrays,” Speech Commun., vol.9, no.5-6, pp.433–442, Dec. 1990.
3. W. Herbdor, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamucu, “Hands-free speech recognition and communication on PDAs using microphone array technology,” Proc. IEEE Automatic Speech Recognition and Understanding Workshop, pp.302–307, Cancun, Mexico, Nov. 2005.
4. L.J. Griffiths and C.W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” IEEE Trans. Antennas Propag., vol.ASSP-30, no.1, pp.27–34, Jan. 1982.
5. S. Gannot, D. Burshtein, and E. Weinstein, “Theoretical analysis of the general transfer function GSC,” Proc. Int. Workshop Acoustic Echo Noise Control, pp.103–106, Darmstadt, Germany, Sept. 2001.
6. S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” IEEE Trans. Signal Process., vol.49, no.8, pp.1614–1626, Aug. 2001.
7. O. Shalvi and E. Weinstein, “System identification using nonstationary signals,” IEEE Trans. Signal Process., vol.44, no.8, pp.2055–2063, Aug. 1996.
8. O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Trans. Signal Process., vol.52, no.7, pp.1830–1847, July 2004.
9. S. Jeong, S. Lee, and M. Hahn, “Dual microphone-based speech enhancement by spectral classification and Wiener filtering,” Electron. Lett., vol.44, no.3, pp.253–254, Jan. 2008.
10. ITU-T Recommendation, “Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders,” ITU-T Recommendation, p.862, Feb. 2001.