euronoise

A two-stage binaural speech enhancement approach for hearing aids with preserving binaural benefits in noisy environments

Junfeng Li[a], Shuichi Sakamoto[b], Satoshi Hongo[c], Masato Akagi[d] and Yôiti Suzuki[b]

[a]Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, 923-1292 Ishikawa, Japan
[b]R.I.E.C., Tohoku University, 2-1, Katahira, Aoba-ku, 980-8577 Sendai, Japan
[c]Faculty of Design and Computer Applications, Miyagi National College of Technology, 48, Nodayama, Medeshima Shiote, 981-1239 Natori, Japan
[d]Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, 923-1292 Sendai, Japan
junfeng@jaist.ac.jp

Previously, we proposed a *Two-Stage BinAural Speech Enhancement* (TS-BASE) approach for hearing aids in adverse environments. In the proposed algorithm, the interfering signal is estimated by cancelling the target signal through an adaptive filter in the first stage and a time-variant Wiener filter is then applied to enhance the target signal given the noisy mixture signals in the second stage. In this paper, we will briefly introduce the proposed TS-BASE algorithm and then focus on the comprehensive experimental evaluations on its speech enhancement performance and its ability of preserving binaural benefits in non-stationary multiple-noise-source environments. Experimental results show that the proposed algorithm outperforms the traditional algorithms in reducing multiple-interfering signals and preserving the ability of sound localization.

1 Introduction

Though persons with normal hearing can understand speech in severely noisy environments, hearing-impaired persons have great difficulty in understanding speech in such conditions due to the hearing loss and the annoying acoustic noise. To facilitate the hearing of hearing-impaired persons, hearing aids have been designed and widely used. One of the main problems for hearing-aid users is the reduction of speech intelligibility in real-life noisy environments. To deal with this problem, efficient speech enhancement techniques have to be integrated into hearing aid signal processing [1].

A large number of speech enhancement algorithms have so far been reported for hearing aids [1]. The *generalized sidelobe canceller* (GSC) was extended to binaural scenarios for hearing aids [2]. In this method, two sub-arrays are independently configured to preserve the *interaural time difference* (ITD) cues in the low frequencies, and a GSC beamformer is applied to all microphone signals with a single output to maximize the spatial directivity in the high frequencies. The main drawback of this method is the low noise reduction performance in the low frequency region. Campbell *et al.* applied a sub-band GSC beamformer to binaural noise reduction for hearing aids [3]. However, a *voice activity detection* (VAD) is needed which does frequently fail, especially in high noise conditions. Moreover, Suzuki *et al.* suggested to introduce binaural cues into the constraints of an adaptive beamformer which realizes the adaptive beamforming and preserves the binaural cues within a certain range of directions [4]. The major problem associated with these algorithms is the low noise reduction performance in multiple-noise-source environments. More recently, Roman *et al.* proposed a speech segregation approach to estimate an ideal *time-frequency* (T-F) binary mask, which finally presents a monaural output signal [5]. Thus, this algorithm loses the spatial benefits resulting from the binaural cues.

Human beings excel at understanding target signal in multi-interference conditions. Inspired by this good ability of human beings, we proposed a *Two-Stage BinAural Speech Enhancement* (TS-BASE) approach that combines interfering signal estimation by cancelling the target signal through adaptive filtering and a consequent stage that controls the transfer function of a time-variant Wiener filter [6]. The proposed TS-BASE system involves no restrictions on interfering signal. In this paper, we will first briefly review the previously proposed TS-BASE system and then concentrate on the experimental evaluations on its noise reduction performance and its ability of preserving the binaural benefits (i.e., sound localization) in non-stationary multiple-noise-source environments. Experimental results confirm the superiorities of the proposed TS-BASE system.

2 Signal Model

For hearing aids in real environments, the microphone signals at the left ear and the right ear do not only differ in the time difference depending on the position of the target sound source to the head, but also in the intensity difference caused by the shadowing effect of the head. Moreover, the microphone signals are also corrupted by the uncorrelated interfering signals. As a result, the observed signals, $X_L(k,\ell)$ and $X_R(k,\ell)$ in the k-th frequency bin and the ℓ-th frame at the left and right ears, can be written as

$$X_L(k,\ell) = S_L(k,\ell) + N_L(k,\ell), \quad (1)$$
$$X_R(k,\ell) = S_R(k,\ell) + N_R(k,\ell), \quad (2)$$

where $S_i(k,\ell) = H_i(k)S(k,\ell)$ and $N_i(k,\ell)$ $(i = L, R)$ are the corresponding *short-time Fourier transforms* (STFTs) of the observed target signal and the uncorrelated interfering signal. $H_i(k)$ denotes the transfer functions between the target sound source to the head, known as *head-related transfer function* (HRTF). Note that the interfering signal here is a combination of multiple noise signals and additional background noise. In this research, the target signal is assumed to be from a certain direction but no restrictions are imposed on the number, location and content of the interfering sources.

3 Binaural noise reduction system

In this section, we give a brief review of our previously proposed binaural noise reduction algorithm [6], which consists of: estimation of interfering signal by cancelling the target signal through an adaptive filter, and a time-variant Wiener filter to enhance the target signal. The block diagram of the proposed system is shown in Fig. 1.

3.1 Estimation of interfering signal

The objective of this part is to estimate interfering signal by cancelling the target signal, as shown in Fig. 1. Note that due to the shadowing effect of the head (e.g., HRTFs), the target signal observed at the left ear is different from that at the right ear. In order to cancel the target signal and produce the interference-only output, we have to compensate this mismatch for the target signal at two ears. To do this, we proposed to exploit two

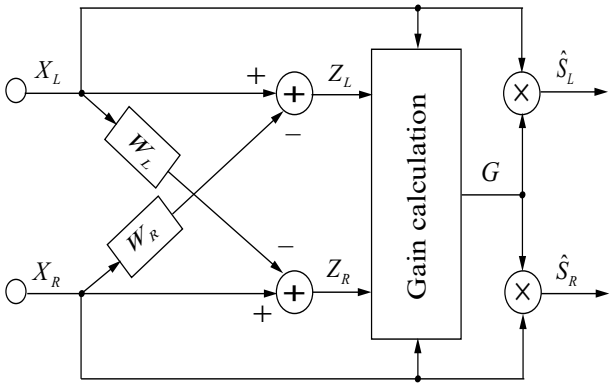

Figure 1: Block diagram of the proposed binaural noise reduction system.

adaptive filters, W_L and W_R, which are used in the left and right channels. In our implementation, the adaptive filters are pre-learned using a white noise sequence of 10 s in the absence of interfering signal. Specifically, given a location of target sound source, the target signals that are used to calibrate the adaptive filters are generated by convoluting the white noise sequence with the corresponding *head-related impulse responses* (HRIRs). With the created target signals as binaural inputs and performing the calibration using the *normalized least mean square* (NLMS), we can obtain two adaptive filters, $\mathbf{W}_L$ and $\mathbf{W}_R$, given by

$$\mathbf{W}_L(\ell+1) = \mathbf{W}_L(\ell)+\mu\frac{\mathbf{X}_L(\ell)}{||\mathbf{X}_L(\ell)||^2}\left[\mathbf{X}_R(\ell)-\mathbf{W}_L^T(\ell)\mathbf{X}_L(\ell)\right], \quad (3)$$

$$\mathbf{W}_R(\ell+1) = \mathbf{W}_R(\ell)+\mu\frac{\mathbf{X}_R(\ell)}{||\mathbf{X}_R(\ell)||^2}\left[\mathbf{X}_L(\ell)-\mathbf{W}_R^T(\ell)\mathbf{X}_R(\ell)\right], \quad (4)$$

where $\mathbf{W}_i(\ell) = [W_i(1,\ell), W_i(2,\ell), \ldots, W_i(K,\ell)]^T$, $\mathbf{X}_i(\ell) = [X_i(1,\ell), X_i(2,\ell), \ldots, X_i(K,\ell)]^T$ $(i = L, R)$, K is the STFT length, and the superscript ^{T} denotes the transpose operator; $\mu = 0.01$ is the step size.

After determining the adaptive filters, their coefficients are fixed and applied to the observed mixture signals in the presence of interfering signal. Since the adaptive filters are learned in the scenarios without interfering signal, the target components of the filter-calibrated left (right) channel inputs should be (approximately, if not exactly) equivalent to the target components of the right (left) channel inputs. Thus, the differential outputs are derived by subtracting the filter-calibrated inputs from the microphone signals, given by

$$\begin{aligned} Z_L(k,\ell) &= X_L(k,\ell) - W_R(k,\ell)X_R(k,\ell) \\ &\approx N_L(k,\ell) - W_R(k,\ell)N_R(k,\ell), \quad (5) \\ Z_R(k,\ell) &= X_R(k,\ell) - W_L(k,\ell)X_L(k,\ell) \\ &\approx N_R(k,\ell) - W_L(k,\ell)N_L(k,\ell). \quad (6) \end{aligned}$$

From Eqs. (5) and (6), we can observe that the target signal has been cancelled, producing the interference-only outputs.

3.2 Enhancement of target signal

For hearing aids, the system that outputs a monaural signal is unacceptable because the noise reduction benefit is consumed by the loss of spatial hearing. To output a binaural signal, the target-cancelled signals, $Z_i(k,\ell)$ derived in the first stage, are used as interference estimate parameters to control the gain function of a speech enhancer for both channels. A real gain function $G(k,\ell)$ is desired in order to minimize distortions from the frequency-domain filter. To do so, we proposed to use a Wiener filter that is the optimal solution for noise reduction in *minimum mean square error* (MMSE) sense. The gain function of the Wiener filter is given by

$$G_{Wiener}(k,\ell) = 1-\frac{Z_L(k,\ell)Z_L^*(k,\ell)+Z_R(k,\ell)Z_R^*(k,\ell)}{X_L(k,\ell)X_L^*(k,\ell)+X_R(k,\ell)X_R^*(k,\ell)}, \quad (7)$$

where the superscript * is the conjugative operator. Note that the target-cancelled signals may have different properties with the interfering components in the observed signals because of the filtering effects introduced by the first stage, however, the target-cancelled signals are highly correlated with the interfering components at the inputs and are still uncorrelated with the target signal, therefore, they still can be used to implement the Wiener filter for reducing interfering signal.

4 Experiments and results

Performance of the proposed binaural noise reduction algorithm was examined in the one- and multiple-noise-source environments, and further compared to that of the traditional algorithms including Roman's system [5], the algorithm in which the *short-time spectral amplitude* (STSA) filter [7] or the *log-spectral amplitude* (LSA) filter [8], instead of Wiener filter used in our proposed algorithm, is used in the second stage for enhancing the target signal shown in Fig. 1.

4.1 Speech enhancement experiments

4.1.1 Experimental configuration

To evaluate the effectiveness of the studied algorithms, two noise acoustic environments, one-noise-source and three-noise-source conditions, were generated. In both environments, ten Japanese sentences were used as target signals and other thirty sentences as interfering signals. In our experiments, the *head-related impulse response* (HRIRs) were obtained from MIT media lab. [9]. The target sound source was placed in the front of the dummy head (i.e., DOA= 0°), and the three interfering sources were located with DOAs of $-60°, 60°, 30°$. The observed signals at two ears were created by convoluting the source (target and interference) signals with the corresponding HRIRs. In the one-noise-source condition, the noisy signals were obtained by summing the observed interfering signals with DOA of 60° and the target signals at different global SNRs $[0, 15]$ dB with the step of 5 dB. In the three-noise-source conditions, the interfering signals at two ears were first generated by mixing the individually observed three interfering signals, and then added to the target signals at the global SNRs same as in the one-noise-source condition.

4.1.2 Speech enhancement results and discussions

To evaluate the studied algorithms in reducing interfering signal and improving speech quality, we selected the objective quality measure known as *perceptual evaluation of speech quality* (PESQ) [10], since it is able to predict subjective quality with good correlation in a very wide range of conditions specified by the ITU-T as recommendation P.862 [10]. The performance was evaluated at the right microphone, the similar tendency was observed for the left microphone inputs.

The experimental results of PESQ averaged across all tested sentences in the one- and three-noise-source conditions are plotted in Fig. 2. The PESQ results in Fig. 2 demonstrate that all studied noise reduction algorithms result in higher PESQ rates, corresponding to higher speech quality of the enhanced signals, compared with the noisy input signals in the one- and three-noise-source conditions at all SNRs. Among the tested algorithms, the proposed algorithm with Wiener filter results in the highest PESQ results in all the tested conditions. With respect to the input noisy signals, the average PESQ improvements achieved by the proposed algorithm amounts to 0.63, and about $0.24, 0.12, 0.10$ compared with Roman's algorithm, the algorithms with STSA filter and LSA filter in the second stage, in the one-noise-source condition. In comparison of the noisy inputs and the traditional algorithms, the PESQ improvements are about $0.49, 0.32, 0.12$ and 0.14, respectively, in the three-noise-source condition. From these results, we note that the performance of the traditional algorithms (especially Roman's algorithm) decreases in the three-noise-source condition. In contrast, our proposed algorithm with Wiener filter demonstrates the only slight performance degradation as the number of microphone increases. As a result, in comparison of the traditional algorithms, our proposed binaural speech enhancement approach is successful in reducing three interfering signals and its performance slightly degrades even if in the non-stationary multiple-noise-source conditions (e.g., the interfering speech used in our experiments).

The proposed algorithm outperforms Roman's algorithm in the sense of noise reduction. This is because that the continuous gain function (i.e., Wiener filter) is exploited in our proposed method, while the discontinuous binary filter is used in Roman's algorithm. Also, the proposed algorithm with Wiener filter gives the higher speech enhancement performance than the ones with STSA and LSA filters in our tested conditions. A possible explanation of this high performance is that Wiener filter could compensate for certain distortions caused by the adaptive filter. However, its mechanism, even if it would exist, is not at all clear at this stage and thus the investigation of the reasons for the high performance of the proposed TS-BASE method with Wiener filter must be an interesting future problem.

4.2 Sound localization experiments

For an effective binaural noise reduction system for hearing aids, in addition to the noise reduction performance, the ability in preserving binaural benefits at the outputs

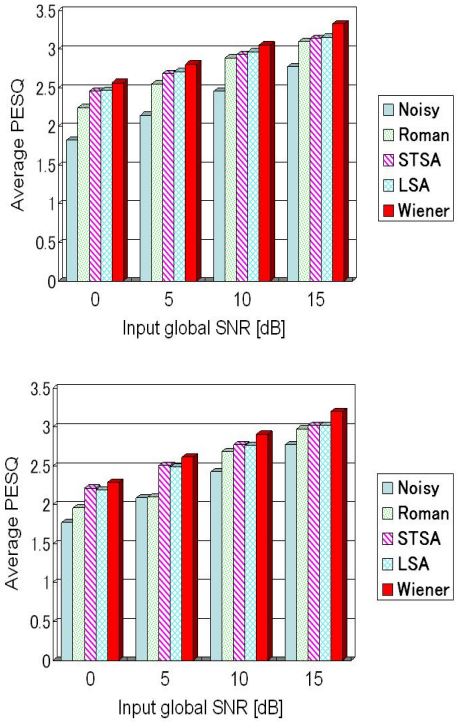

Figure 2: PESQ results in the one-noise-source condition (upper) and in the three-noise-source condition (lower).

is another important characteristic. To examine this ability of our proposed algorithm, we conducted subjective sound localization experiments in the one- and three-noise-source environments.

4.2.1 Experimental configuration

In the localization experiments, the target sound source moved from $-90°$ to $90°$ in both one- and three-noise-source conditions. The one-noise-source conditions involved one interfering signal with DOA of $0°$, the three-noise-source conditions included three interferences with DOAs of $-60°, 0°$ and $30°$. The observed mixture signals were generated by adding the interfering signals into the target signals at the SNR of 0 dB, and processed by our proposed algorithm with Wiener filter which gives the highest PESQ results as shown in Section 4.1.2. The resultant enhanced signals were then randomly presented to six volunteers with normal-hearing ability through a headphone. Each listener was firstly pre-trained using the clean signals given the "real" DOAs in the absence of interfering signals. After that, the listeners attended the testing procedure in which the enhanced target signals were presented, and were then instructed to give the perceived directions of the enhanced signals.

4.2.2 Sound localization results and discussions

The localization results in the one- and three-noise-source conditions are plotted in Fig. 3 as well as the range of $\pm 15°$ relative to the "real" DOAs of input signals. Fig. 3 shows that the perceived directions are the same as or very close to the "real" directions in the one- and three-noise-source conditions. Further observations on

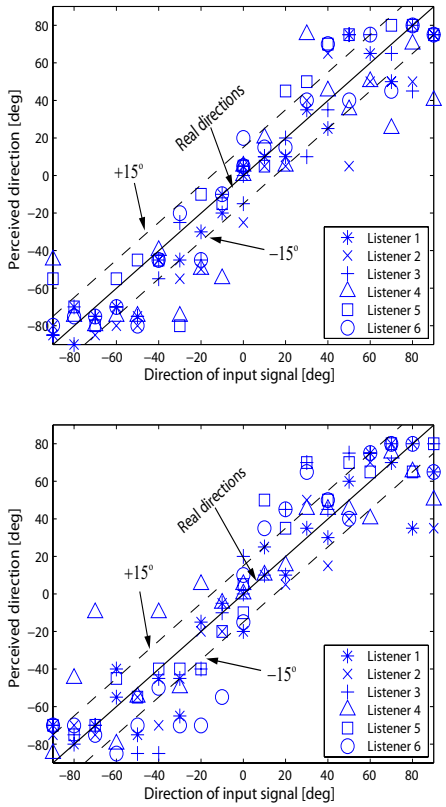

Figure 3: Results of the perceived directions of the signals processed by the proposed algorithm against the directions of input signals in the one-noise-source condition (upper) and the three-noise-source condition (lower) at the SNR of 0 dB.

the localization results for the input signals with different DOAs, we note that the perceived direction for the enhanced signal is much closer to the "real" DOA if the input DOA is small (e.g., $-40° < \text{DOA} < 40°$). Moreover, we observe the variances of the perceived DOAs in the three-noise-source condition are also larger than those in the one-noise-source condition, which is due to the relatively higher speech quality of the enhanced signals in the one-noise-source condition. On the whole, the localization results shown in Fig. 3 proved that the proposed algorithm is able to preserve the binaural benefits at the outputs.

In the proposed binaural noise reduction algorithm, one real gain was calculated and shared in both left and right channels. This mechanism preserves the binaural (e.g., ITD and ILD) cues at the outputs to a certain degree. Note that the preserved binaural cues are not those of the target signal, but should approximate. This is because the interfering signals are greatly suppressed by the proposed algorithm, which further markedly decrease the effects of the interfering signals on the binaural cues of the target signal. As a result, the proposed algorithm is able to localize the target sound source based on the enhanced binaural signals, which is mainly benefited from the preserved binaural cues.

5 Conclusion

In this paper, we first briefly introduced our previously proposed *Two-Stage BinAural Speech Enhancement* (TSBASE) method, which consists of: interference estimation through an adaptive filter and speech enhancement through a Wiener filter. Our concentration was then paid to the experimental evaluations on the proposed algorithm in the sense of speech enhancement performance with the PESQ measure and the ability of preserving the binaural benefits at the outputs with sound localization. Experimental results confirmed its effectiveness in both speech enhancement and preservation of binaural benefits.

6 Acknowledgments

This study was supported by Sendai Intelligent Knowledge Cluster and Grant-in-Aid for Young Scientists (B) (No. 19700156) from the Ministry of Education, Science, Sports and Culture of Japan.

7 References

References

[1] V. Hamacher, et al., "Signal processing in high-end hearing aids: State of the art, Challenges, and Future Trends," *EURASIP Journal on ASP*, vol. 18, pp. 2915-2929, 2005.

[2] J.G. Desloge, W.M. Rabinowitz and P.M. Zurek, "Microphone-array hearing aids with binaural output-part I: Fixed-processing systems," *IEEE Trans. SAP*, vol. 5, pp. 529-542, 1997.

[3] D. Campbell and P. Shields, "Speech enhancement using sub-band adaptive Griffiths-Jim signel processing," *Speech Comm.*, vol. 39, pp. 97-110, 2003.

[4] Y. Suzuki, S. Tsukui, F. Asano, R. Nishimura and T. Sone, "New design method of a binaural microphone array using multiple constraints," *IEICE Trans. Fundamentals*, vol. E82-A, pp. 588-595, 1999.

[5] N. Roman, S. Srinivasan and D. Wang, "Binaural segregation in multisource reverberant environments," *JASA*, vol. 120, no. 6, pp. 4040-4050, 2006.

[6] J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, "A speech segregation approach for binaural hearing aids," in *Proc. the 22nd Signal Processing Symposium*, pp. 263-268, Sendai, Japan, Nov. 2007.

[7] Y. Ephraim, D. Malah,"Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 32, pp. 1109-1121, 1984.

[8] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. ASSP*, vol. 33, pp. 443-445, 1985.

[9] http://sound.media.mit.edu/KEMAR.html.

[10] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of 3.1 kHz handset telephony (narrow-band) networks and speech codecs," Feb. 2001.