TWO-STAGE BINAURAL SPEECH ENHANCEMENT WITH WIENER FILTER BASED ON EQUALIZATION-CANCELLATION MODEL

Junfeng Li[1], *Shuichi Sakamoto*[2], *Satoshi Hongo*[3], *Masato Akagi*[1] *and Yôiti Suzuki*[2]

[1] School of Information Science, Japan Advanced Institute of Science and Technology
[2] Research Institute of Electrical Communication, Tohoku University
[3] Department of Design and Computer Application, Miyagi National College of Technology

ABSTRACT

The equalization-cancellation (EC) model has been extensively studied for expressing binaural masking level difference (BMLD) in psychoacoustics. Few research focuses on applying this psychoacoustic model to speech processing applications, such as speech enhancement. In this paper, we propose a two-stage binaural speech enhancement with Wiener filter (TS-BASE/WF) based on the EC model. In this proposed TS-BASE/WF, interfering signals are first estimated by equalizing and cancelling the target signal based on the EC model, and a time-variant Wiener filter is then applied to enhance the target signal given noisy mixture signals. The main advantages of the proposed TS-BASE/WF are: (1) effectiveness in dealing with non-stationary multiple-source interfering signals; (2) success in localizing the target sound source after processing. These advantages were confirmed by comprehensive experiments in different spatial scenarios in terms of speech enhancement and sound localization.

Index Terms— Equalization-cancellation (EC) model, TS-BASE/WF, Speech enhancement, Sound source localizaton.

1. INTRODUCTION

The last decades have witnessed significant advancements in speech signal processing and in binaural hearing in psychoacoustics, usually in a separative way. Speech signal processing has activated the rapid progress in speech applications, e.g., speech enhancement. Meanwhile, psychoacoustic research in binaural hearing shows that additional great benefits in understanding a signal in noise could be obtained if the speech and noise come from different directions. Moreover, the binaural cues in signals also make it possible to localize their sources and give birth to perceptual impression on the acoustical scene in realistic environments. Therefore, great interest has recently been paid to develop binaural speech enhancement systems based on the knowledge of psychoacoustics and signal processing.

In speech enhancement, two-microphone noise reduction has been extensively researched because of its simplicity in implementation and its spatial filtering ability [1, 2, 3, 4]. Dorbecker *et al.* proposed to extend the single-channel spectral subtraction to the binaural scenario based on the assumption of zero correlation between the noise signals on two microphones [1], which is not satisfied in practical environments. Kollmeier *et al.* introduced a binaural noise reduction scheme based on the interaural phase difference (IPD) and interaural level difference (ILD) cues in the frequency domain [2]. This method was further considered by Nakashima *et al.*, named as frequency domain binaural model (FDBM), by discriminating the target and interfering signals based on the estimates of their directions [3], which is however quite difficult in real conditions. Lotter *et al.* proposed a dual-channel speech enhancement based on superdirective beamforming under the assumption of a diffuse noise field [4]. Moreover, Klasen *et al.* extended the monaural multi-channel Wiener filtering (MWF) [5] to the binaural scenario to preserve the binaural cues. However, the adaptive MWF beamformer with two microphones is only optimal for cancelling a single directional interference. A similar problem is also associated with blind source separation (BSS)-based binaural systems, e.g., the system proposed by Aichner *et al.* [6].

In psychoacoustics, binaural masking level difference (BMLD) is a psychoacoustic effect whereby the detection of a signal in noise is improved when either the phase or level differences of the signal at two ears are not the same as those of the maskers. To account for the BMLD effect, many binaural models have been presented, including the interaural cross-correlation-based model [7], and the equalization-cancellation (EC) model that is based on the cancellation of binaural maskers [8]. The cross-correlation model can interpret the BMLD effect by essentially utilizing the similarities of binaural inputs. On the other hand, the EC model expresses the BMLD effect based on the dissimilarity of binaural inputs.

Based on the psychoacoustic EC model, in this paper, we propose a two-stage binaural speech enhancement approach with Wiener filter (TS-BASE/WF) for high-quality realistic speech communication. The proposed TS-BASE/WF first estimates the interfering signals by performing the equalization and cancellation processes for the target signal based on the EC model, and then enhances the target signal by using a Wiener filter. Experimental results show that the proposed TS-BASE/WF is able to suppress non-stationary multiple interference signals and to localize the target signal after processing in different spatial conditions.

2. THE EQUALIZATION-CANCELLATION MODEL

The equalization-cancellation (EC) model was originally developed by Durlach [8] and further improved by Culling and Summerfiled [9]. In the original EC model, it was assumed that the auditory system transforms the signals arriving at two ears so that the masker components are "equalized" (the E process), and then subtracts the total signal in one ear from the total signal in the other ear (the C process) [8]. This model was recently improved in [9], where the E and C processes were independently performed for the interfering signal in each channel. Although these EC mod-

This research is partially supported by the SCOPE (071705001) of Ministry of Internal Affairs and Communication (MIC), Japan.

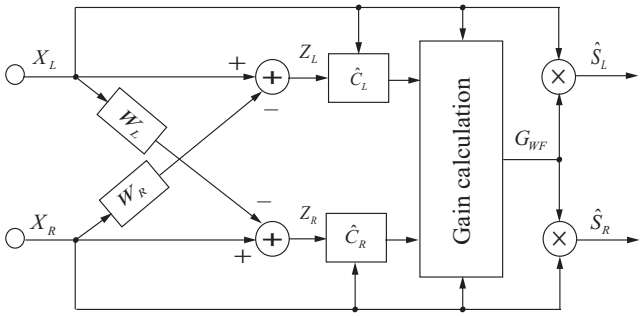

Figure 1: Block diagram of the proposed TS-BASE/WF algorithm.

els could explain many psychoacoustic effects (e.g., BMLD), they function well only in single interference conditions [8, 9].

3. TWO-STAGE BINAURAL SPEECH ENHANCEMENT WITH WIENER FILTER

Based on the EC model, a two-stage binaural speech enhancement approach with Wiener filter (TS-BASE/WF) is developed, which consists of: (1) interferences estimation by equalizing and cancelling the target signal components, followed by a compensation process; (2) target signal enhancement by a Wiener filter. The block diagram of the proposed system is shown in Fig. 1.

3.1. Signal model

For binaural applications in noisy conditions, the observed signals, $X_L(k,\ell)$ and $X_R(k,\ell)$, in the kth frequency bin and the ℓth frame at the left and right ears, are written as

$$X_i(k,\ell) \quad = \quad S_i(k,\ell) + N_i(k,\ell), \;\; i = L, R, \qquad (1)$$

where $S_i(k,\ell) = H_i(k,\ell)S(k,\ell)$ and $N_i(k,\ell)$ are respectively the spectra of the target and interfering signals; $H_i(k)$ represents the transfer functions between the target sound source to two ears, referred to as *head-related transfer function* (HRTF) in the context of binaural hearing. Note that the interfering signals, $N_i(k,\ell)$, might be a combination of multiple interfering signals and background noise. In this research, the direction of the target signal is known a priori; but no restrictions are imposed on the number, location and content of the interfering noise sources.

3.2. Estimation of interfering signals

3.1.1 Equalization and Cancellation of the target signal

In binaural applications, HRTFs are normally involved to include the shadowing effects of the head which exhibits the differences in amplitude, phase and onset time for the signals at the left and right ears. The cancellation of the target signal is achieved on the basis of the EC model, yielding the interference-only outputs. It is specifically realized in the following two steps.

1. *In the "equalization" (E) process*, two adaptive filters are applied to the left and right input signals for equalizing the target signal components in these inputs. Given the binaural inputs, two equalizers, $W_L(k,\ell)$ and $W_R(k,\ell)$, can be obtained by using the normalized least mean square (NLMS) algorithm, given by

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

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

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 transposition operator; μ is the step size.

Based on the assumption that the direction of the target signal is known a priori, in this research, the two equalizers are pre-learned in the absence of interfering signals. Specifically, the binaural input signals generated by convolving a white noise sequence of 10 s duration with the corresponding head-related impulse response (HRIR) are used as inputs of the NLMS algorithm to calibrate the two equalizers.

2. *In the "cancellation" (C) process*, the coefficients of two equalizers are fixed and applied to the observed mixture signals in the presence of interfering signals. Since the equalizers are calibrated in the scenarios without interfering signals, the target components of the filter-calibrated left (right) channel input should be approximately, if not exactly, equivalent to the target components of the right (left) channel input. As a result, the target-cancelled signals are derived by subtracting the filter-calibrated inputs at one ear from the input signals at the other ear, given by

$$\begin{aligned} Z_L(k,\ell) &= X_L(k,\ell) - W_R(k,\ell)X_R(k,\ell) \\ &\approx N_L(k,\ell) - W_R(k,\ell)N_R(k,\ell), \quad (4) \\ Z_R(k,\ell) &= X_R(k,\ell) - W_L(k,\ell)X_L(k,\ell) \\ &\approx N_R(k,\ell) - W_L(k,\ell)N_L(k,\ell). \quad (5) \end{aligned}$$

From Eqs. (4) and (5), we observe that the target signal has been cancelled, yielding the interference-only outputs.

The original EC model and its recent variants perform the E and C processes for the interfering signals, and enable to successfully reduce only one directional interfering signal with two microphones. Thus, they cannot function well in multiple-source and diffuse noise environments. In contrast, the EC processes realized in the proposed TS-BASE/WF system are intended to equalize and cancel the target components in the signals at two ears, which yields the interference-only outputs that might include the energy of multiple interfering signals and diffuse noise. Thus, this realization of the EC model in the TS-BASE/WF system can be used to further address the problem of multiple interfering signals in adverse environments.

3.1.2 Compensation of interfering signal estimates

As mentioned in the last subsection, the purpose of the EC processes in our proposed TS-BASE system is to estimate the interfering components by equalizing and cancelling the target signal components. Note that although the EC processes have successfully cancelled the target components, as shown in Eqs. (4) and (5), the target-cancelled outputs are different from the interference components in the input mixture signals because of the filtering effects introduced by the two equalizers.

To address this issue, we propose to exploit a time-variant frequency-dependent compensation factor, $C_i(k,\ell)$, for mapping the target-cancelled signals to the interfering components in the input mixture signals. This compensation factor $C_i(k,\ell)$ is derived by minimizing the mean square error between the target-cancelled signal and the input mixture signal under the assumption of zero correlation between the target signal and interfering signals, formulated as

$$\hat{C}_i(k,\ell) = \arg\min_{C_i} E\big[X_i(k,\ell) - Z_i(k,\ell)C_i(k,\ell)\big],\ i = L, R \quad (6)$$

where E is the expectation operator. The optimal compensation factor can be found by setting the derivative of the cost function with respect to the factor $C_i(k,\ell)$ to zero. Based on Wiener theory, the optimal compensator, $C_i^{\text{opt}}(k,\ell)$, is given by

$$C_i^{\text{opt}}(k,\ell) = \frac{\phi_{X_i Z_i}(k,\ell)}{\phi_{Z_i Z_i}(k,\ell)}, \qquad i = L, R \quad (7)$$

where $\phi_{X_i Z_i}(k,\ell)$ denotes the cross-correlation spectrum of $X_i(k,\ell)$ and $Z_i(k,\ell)$; and $\phi_{Z_i Z_i}(k,\ell)$ is the auto-correlation spectrum of $Z_i(k,\ell)$.

3.3. Enhancement of target signal

For binaural applications, the system that outputs binaural signals is much preferred. In the proposed TS-BASE/WF system, the compensated interference estimates are used to control the gain function of a speech enhancer which is shared in both channels for binaural cue preservation. In this research, the improved Wiener filter based on the *a priori* SNR is adopted, due to the simplicity in its implementation and its ability in reducing "musical noise", formulated as [10]

$$G_{WF}(k,\ell) = \frac{\xi(k,\ell)}{1+\xi(k,\ell)}, \quad (8)$$

where $\xi(k,\ell)$ is the *a priori* SNR calculated as (k and ℓ are omitted for simplicity):

$$\xi = \frac{E\Big[S_L S_L^* + S_R S_R^*\Big]}{E\Big[\big(C_L Z_L\big)\big(C_L Z_L\big)^* + \big(C_R Z_R\big)\big(C_R Z_R\big)^*\Big]}, \quad (9)$$

where the superscript * is the conjugation operator. The estimate of the *a priori* SNR, $\xi(k,\ell)$, is updated in a decision-directed scheme that significantly decreases the residual "musical noise".

4. EXPERIMENTS AND RESULTS

The performance of the proposed TS-BASE/WF system was examined in one- and multiple-noise-source conditions, and further compared to that of the traditional algorithms including the two-channel spectral subtraction (TwoChSS) [1], the frequency-domain binaural model (FDBM) [3], and the two-channel superdirective beamformer (TwoChSDBF) [4]. A large number of experiments were carried out to comprehensively evaluate the performance of the tested algorithms, with respect to speech enhancement and sound localization, in various spatial configurations and in terms of objective and subjective evaluation measures. Due to the space limitation, only the subjective evaluations and results are presented in this paper, and more experimental results are described in [11].

4.1. Speech enhancement experiments

4.1.1 Experimental configuration

In subjective speech enhancement evaluations, 6 utterances were selected from the NTT database and used as the target speech signals, and 24 other different utterances as the interfering signals. The observed mixture signals at two ears were generated by convolving the "dry" (target and interference) signals with the HRIRs obtained from MIT media lab. The evaluations were performed at the SNR of 0 dB in the following spatial configurations: S_0N_{60}, $\text{S}_0\text{N}_{90,180,270}$, $\text{S}_0\text{N}_{60,120,180,270}$ and S_{90}N_0, where S_xN_y denotes a spatial scenario with a target signal (S) arriving from the direction x°, and one or multiple noise sources (N) arriving from the direction(s) y°.

The resulting 24 (4×6) noisy speech utterances were then processed by 4 tested algorithms. The processed 96 (24×4) speech signals, along with the 24 unprocessed signals as reference, were then randomly presented to ten graduate students with normal hearing ability through a headphone at a comfortable volume in a soundproof room. Each listener was instructed to rate the speech quality based on their preference in terms of *mean opinion score* (MOS). To examine the speech enhancement performance of the tested algorithms, the MOS improvement ΔMOS achieved by each algorithm was calculated as $\Delta\text{MOS} = \text{MOS}_{enhanced} - \text{MOS}_{unproc}$, where MOS_{unproc} and $\text{MOS}_{enhanced}$ are the MOS scores of the unprocessed signal and the enhanced signal obtained with the tested algorithm.

4.1.2 Speech enhancement results and discussion

The improvements in the MOS scores of the studied algorithms in different acoustic scenarios are plotted in Fig. 2. It can be seen that all tested algorithms yield different degrees of MOS improvements at two ears in all tested conditions.

In the conditions in which the target signals arrive from 0°, only small MOS improvements with the TwoChSDBF algorithm were observed in our tests. This low speech enhancement ability is attributed to the assumption of a diffuse noise field in its design, which fails in the tested conditions. In comparison with the TwoChSDBF algorithm, the TwoChSS algorithm provides much larger MOS improvements in these conditions. Based on the interaural information of the binaural inputs, the FDBM algorithm shows relatively robust MOS improvements as the number of interfering signal increases. In contrast, the proposed TS-BASE/WF algorithm yields the largest MOS improvements, i.e., the highest speech quality, amongst the tested algorithms in all spatial configurations, and its performance in ΔMOS shows only a slight decrease at two ears with an increasing number of interfering signals.

More importantly, in the acoustic condition S_{90}N_0, the traditional TwoChSS method does not function well, since it normally assumes that the target signal comes from 0°. The limited ability of the TwoChSDBF algorithm is attributed to its unreasonable noise field assumption in the tested condition. The low performance in ΔMOS of the FDBM is due to its failure in discriminating the target and interfering signals based on the binaural cues. In contrast, the proposed TS-BASE/WF yields the very large MOS improvement in this condition.

4.2. Evaluations for sound source localization

4.2.1 Experimental configuration

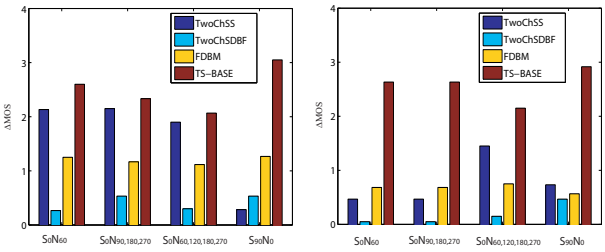

Figure 2: MOS improvements of the studied algorithms at the SNR=0dB at the left (a) and right (b) ears in the different acoustical conditions.

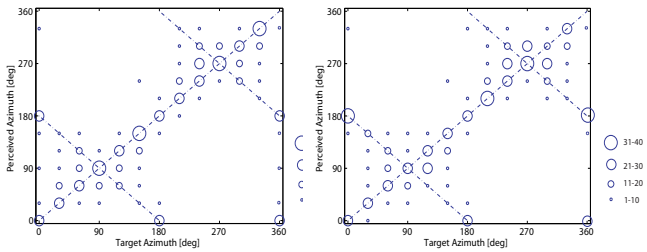

Figure 3: Results of sound localization tests in the one-noise-source condition S_xN_0 (a), and in the three-noise-source condition $S_xN_{90,180,270}$ (b), where $0^\circ \leq x \leq 360^\circ$.

The objective evaluations in [11] demonstrated that the proposed TS-BASE/WF is able to markedly decrease the ITD and ILD errors compared with the traditionally tested algorithms, therefore, in this subsection, only the proposed TS-BASE algorithm was evaluated to further subjectively confirm its ability in sound localization via listening tests. In the evaluations, the same target and interfering signals were used as those in speech enhancement experiments. The binaural signals were then generated by convolving these signals with the corresponding HRIRs to generate the following spatial conditions: (1) the one-noise-source condition ($S_{0:30:360}N_0$); (2) the three-noise-source conditions ($S_{0:30:360}N_{90,180,270}$), where the target sound source moves from 0° to 360°. The observed mixture signals were generated by adding the interfering signals into the target signals at the SNR of 0 dB, and then processed by the TS-BASE/WF algorithm. The resultant enhanced signals were then randomly presented to the listeners who also participated in the speech enhancement experiments through headphones in the soundproof room. Each listener was firstly pre-trained using the binaural 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 randomly presented, and were then instructed to give the perceived directions of the enhanced signals.

4.2.2 Localization results and discussion

The localization results in the one- and three-noise-source conditions are plotted in Fig. 3. The diameter of each circle is proportional to the number of responses. The ordinate of each panel is the perceived direction, and the abscissa is the target direction. Fig. 3 shows that the responses are distributed along a diagonal line, that is, the perceived directions closely agree with the "real" target directions. Further observation illustrates that when the target signal lies in the front and rear regions (0° and 180°), most subjects are able to perceive the correct target directions in both spatial scenarios; while in the lateral area (90° and 270°), the perceived directions are dispersed around the target directions. More importantly, the front-back confusion was evidently observed in both the one- and three-noise-source conditions. In comparison to the results in these two spatial conditions, the variances of the perceived directions for the target signals in the one-noise-source conditions are slightly lower than those in the three-noise-source conditions. As a result, the proposed TS-BASE/WF algorithm is able to successfully localize the target signal in the complex acoustical environments.

5. CONCLUSION

In this paper, we proposed a two-stage binaural speech enhancement with Wiener filter algorithm (TS-BASE/WF) based on the psychoacoustic equalization-cancellation (EC) model. In the TS-BASE/WF, the interfering signal is first estimated by equalizing and cancelling the target signal through adaptive filtering based on the EC model, followed by a compensating process, and target signal enhancement by the time-variant Wiener filter. Subjective evaluations in various spatial conditions indicate that the proposed TS-BASE/WF algorithm yields the highest MOS improvements (i.e., the highest speech quality) and a high ability in accurately localizing the target sound source.

6. REFERENCES

[1] M. Dorbecker, S. Ernst, "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation," in *Proc. EUSIPCO*, pp. psp-9, 1996.

[2] B. Kollmeier, J. Peissig, V. Hohmann, "Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain," *Scand. Audiol. Suppl.* vol. 38, pp. 28-38, 1993.

[3] H. Nakashima, *et al.*, "Frequency domain binaural model based on interaural phase and level differences," *Acoust. Sci. & Tech.*, vol. 24, no. 4, pp. 172-178, 2003.

[4] T. Lotter, B. Sauert and Peter Vary, "A stereo input-output superdirective beamformer for dual channel noise reduction," in *Proc., Eurospeech*, pp. 2285-2288, 2005.

[5] T.J. Klasen, *et al.*, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues processing," *IEEE Trans. on Signal Processing*, vol. 55, no. 4, pp. 1579-1585, 2007.

[6] R. Aichner, *et al.*, "Multi-channel source separation preserving spatial information," in *Proc. ICASSP*, pp. I.5-8, 2007.

[7] L. A. Jeffress, "A place theory of sound localization," *J. Comparative and Physiological Psychology*, vol. 41, 35-39, 1948.

[8] N.I. Durlach, "Equalization and cancellation theory of binaural masking level differences," *JASA*, vol. 35, no. 8, pp. 1206-1218, 1963.

[9] J.F. Culling and Q. Summerfield, "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," *JASA*, vol. 98, pp. 785-797, 1995.

[10] P. Scalart, J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, vol. 2, pp. 629632, 1996.

[11] J. Li, *et al.*, "Objective evaluations of two-stage binaural speech enhancement with Wiener filter for speech enhancement and sound localization," *To appear in Proc. Int. Symposium on Auditory and Audiological Research*, Denmark, August, 2009.