Objective evaluations of two-stage binaural speech enhancement with Wiener filter for speech enhancement and sound localization

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, Japan*

[2] *Research Institute of Electrical Communication, Tohoku University, Japan*

[3] *School of Information Science, Miyagi National College of Technology, Japan*

For high-quality speech communication, we previously proposed a two-stage binaural speech enhancement with Wiener filter (TS-BASE/WF) approach inspired by the equalization-cancellation (EC) theory, to suppress interfering signals and preserve impression of acoustic scene. In the proposed TS-BASE/WF, the interfering signal is first estimated by equalizing and cancelling the target signal through two equalizers and a time-variant Wiener filter is then applied to enhance the target signal given the noisy mixture signals. In this paper, we pay main attention to the comprehensive experimental evaluations on its speech-enhancement performance and its ability in preserving binaural benefits in a variety of acoustic conditions. Experimental results show that the TS-BASE/WF approach is able to suppress non-stationary multiple interfering signals and enhance the target signal which is expected to improve the quality of speech communication, and succeeds in preserving the binaural cues which is expected to give birth to the perceptual impression of the auditory scene, in all tested spatial scenarios.

INTRODUCTION

The last decades have witnessed significant advancements in speech signal processing and in binaural hearing in psychoacoustics, usually in a separate way. Speech signal processing has activated the rapid progress in speech applications, e.g., speech enhancement. Meanwhile, the psychoacoustic researches in binaural hearing show 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 the perceptual impression of the acoustical scene in realistic environments. Therefore, great interests have been recently paid to develop binaural speech enhancement systems based on the knowledge of psychoacoustics and signal processing.

In recent years, two-microphone noise reductions have been extensively researched because of its simplicity in implementation and its spatial filtering ability (Dorbecker and Ernst, 1996; Kollmeier *et al*., 1993; Nakashima *et al*., 2003; Lotter *et al*., 2005). Dorbecker and Ernst (1996) proposed to extend the single-channel spectral subtraction

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

to the binaural scenario based on the assumption of zero correlation between the noise signals on two microphones (Dorbecker and Ernst, 1996), which is not satisfied in practical environments. Kollmeier *et al.* (1993) introduced a binaural noise reduction scheme based on the interaural phase difference (IPD) and interaural level difference (ILD) cues in the frequency domain (Kollmeier *et al.*, 1993). This method was further considered by Nakashima *et al.* (2003), named as frequency domain binaural model (FDBM), by discriminating the target and interfering signals based on the estimates of their directions, which is however quite difficult in real conditions. Lotter *et al.* (2005) proposed a dual-channel speech enhancement based on superdirective beamforming under the assumption of a diffuse noise field. Moreover, Klasen *et al.* (2007) extended the monuaral multi-channel Wiener filtering (MWF) to the binaural scenario to preserve the binaural. However, the adaptive MWF beamformer with two microphones is only optimal for cancelling a single directional interference. The similar problem is also associated with blind source separation (BSS) based binaural systems, e.g., the one proposed by Aichner *et al.* (2007).

More recently, inspired by the equalization-cancellation (EC) theory that accounts for the binaural masking level difference (BMLD) in psychoacoustics, we proposed a two-stage binaural speech enhancement with Wiener filter (TS-BASE/WF) approach for high-quality realistic speech communication (Li *et al.*, 2009). In the proposed TS-BASE/WF, the interfering signals is estimated by performing the equalization and cancellation processes for the target signal inspired by the EC theory, and the target signal is enhanced by using a Wiener filter. In this paper, we first briefly review the proposed TS-BASE/WF algorithm, and then focus on its performance evaluation with regard to speech enhancement and sound localization. The effectiveness of the TS-BASE/WF algorithm in suppressing multiple interference signals is assessed by the objective signal to noise ratio (SNR) improvement, and its ability in preserving the binaural cues is examined through the objective evaluation using binaural cue errors.

TWO-STAGE BINAURAL SPEECH ENHANCMENT WITH WIENER FILTER

Inspired by the EC model, the two-stage binaural speech enhancement approach with Wiener filter (TS-BASE/WF) was recently developed (Li *et al.*, 2009). The TS-BASE/WF 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.

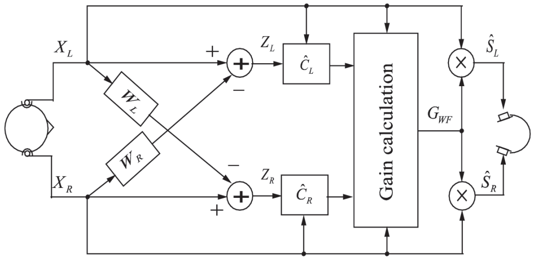

Fig. 1: Block diagram of TS-BASE/WF.

Estimation of interference signal

In binaural applications, head related transfer functions (HRTFs) are normally involved to include the shadowing effects of the head. The cancellation of the target signal is achieved through the equalization and cancellation procedures, yielding the interference-only outputs. It is realized in the following steps:

(1) In the "equalization" process, two 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 can be obtained by using a normalized least mean square (NLMS) algorithm. Based on the assumption that the direction of the target signal is known *a priori*, two equalizers are pre-learned in the absence of interference signals.

(2) In the "cancellation" process, the coefficients of two equalizers are fixed and applied to the observed mixture signals in the presence of interference signals. The target-cancelled signals are derived by subtracting the filter-calibrated inputs at one ear from the input signals at the other ear.

(3) In the "compensation" process, a time-variant frequency-dependent compensation factor is exploited to mapping the target-cancelled signals to the interference components in the input mixture signals. This compensation factor is derived by minimizing the mean square error (MMSE) between the target-cancelled signal and the input mixture signal under the assumption of zero correlation between the target signal and interference signals.

Enhancement of target signal

For binaural applications, the system that outputs binaural signals is much preferred. In the proposed TS-BASE/WF, 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. Specifically, a Wiener filter is used because it is the optimal solution for noise reduction in MMSE sense. Its real gain function contributes to minimize the speech distortion from the frequency-domain filter. The decision-directed adaption mechanism of the *a priori* SNR helps to reduce the "musical noise" and improve speech quality.

OBJECTIVE EVALUATIONS OF TS-BASE/WF ALGORITHM

Performance of the proposed TS-BASE/WF 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) (Dorbecker and Ernst, 1996), the frequency-domain binaural model (FDBM) (Nakashima *et al*., 2003), and the two-channel superdirective beamformer (TwoChSDBF) (Lotter *et al*., 2005). Numerous experiments were carried out to assess the performance of the tested algorithms with regard to speech enhancement and binaural cue preservation in various spatial configurations. The subjective evaluations are presented in (Li *et al*., 2009).

Objective evaluations for speech enhancement

In speech enhancement experiments, 50 continuous speech sentences uttered by three male and two female speakers were randomly selected from a NTT database that has a sampling rate of 44.1 kHz at 16 bit resolution. Among these utterances, 10 sentences were used as the target speech signals, the other 40 were used as the interference signals. These signals were then convolved with head related impulse responses (HRIRs) measured at the MIT Media Laboratory to generate the binaural target and interference signals. The binaural target and interference signals were down-sampled to 8 kHz. The interference signals were then scaled to obtain an average input SNR of 0 dB across two channels before being added to the target signals. The binaural noisy input signals were finally generated by adding the scaled binaural interference signals to the binaural target signals.

To examine the efficacy of the studied systems, we performed evaluations in various spatial configurations as listed in Table 1. $S_\theta N_\varphi$ denotes the spatial scenario in which the target signal (S) arrives from the direction θ and interference signal(s) (N) come from direction(s) φ. Directions are defined clockwise with 0° being directly in front of the listener.

Scenario	Spatial Scenarios	Description
One-noise-source	S_0N_φ	Speech source at 0°; φ between 0 ° and 330°
	$S_{45}N_{315}$	Speech source at 45°; noise source at 315 °
	$S_{90}N_0$	Speech source at 90°; noise source at 0 °
	$S_{90}N_{270}$	Speech source at 90°; noise source at 270 °
Two-noise-source	S_0N_{2a}	Noise sources at 60°, 300°
	S_0N_{2b}	Noise sources at 120°, 240°
	S_0N_{2c}	Noise sources at 90°, 270°
Three-noise-source	S_0N_{3a}	Noise sources at 90°, 180°, 270°
	S_0N_{3b}	Noise sources at 30°, 60°, 300°
Four-noise-source	S_0N_{4a}	Noise sources at 60°, 120°, 180°, 270°
	S_0N_{4b}	Noise sources at 45°, 135°, 225°, 315°

Table 1: List of spatial scenarios, $S_\theta N_\varphi$, under which the speech enhancement capability of the studied algorithms were evaluated.

Evaluation measure

The improvement in SNR was used to evaluate the speech enhancement performance of the proposed TS-BASE/WF and traditional algorithms objectively. It is defined as

$$\Delta \mathrm{SNR} = \mathrm{SNR}_o - \mathrm{SNR}_i \quad , \qquad \text{(Eq. 1)}$$

where SNR_o and SNR_i are the SNRs of the output enhanced signal and the input noisy signal. The SNR is defined as the ratio of the power of clean speech to that of noise signal embedded in the noisy input signal (SNR_i) or the enhanced signal by the studied algorithms (SNR_o). A higher ΔSNR means a higher improvement in speech quality by speech enhancement processing techniques.

Evaluation results

The ΔSNR results in the one-noise-source conditions presented in Fig. 2 show that all tested algorithms produce positive ΔSNRs (i.e. improved speech quality), and that these ΔSNRs vary greatly with the incoming direction of the interference signal. The TwoChSDBF and FDBM algorithms yield low ΔSNRs under all tested conditions. Compared with the TwoChSDBF and FDBM algorithms, the TwoChSS algorithm yields much larger ΔSNRs. In contrast with all traditional algorithms, the proposed TS-BASE/WF algorithm provides the highest ΔSNRs in all tested conditions, especially when the interference signal is close to the ear under evaluation. The high speech-enhancement performance of the proposed TS-BASE/WF results from its accurate noise estimation capability through the equalization and cancellation processes for the target signal. All tested algorithms fail to distinguish the target signal and interference signals based on their binaural cues. Similar results are observed for the right ear.

The ΔSNR results shown in Fig. 3 demonstrate that the studied algorithms can enhance the speech quality (i.e. the positive ΔSNR) at the left and right ears in all multiple-noise-source conditions. The TwoChSDBF algorithm gives the lowest SNR improvements. Comparatively, the FDBM and TwoChSS systems coequally produce much larger ΔSNRs. The TS-BASE/WF algorithm provides significant SNR improvements at both left and right ears in the presence of multiple interference sources. Another important observation is that in the conditions with non-zero arrival direction of the target signal, the traditional TwoChSDBF and TwoChSS algorithms show very limited SNR improvements. The FDBM approach gives much higher SNR improvements. Regarding results observed at the right ear, the TwoChSS and FDBM algorithms show the markedly decreased ΔSNR in the $S_{90}N_0$ scenario and even the negative ΔSNRs in $S_{90}N_{270}$ and $S_{45}N_{315}$ conditions, and the TwoChSDBF algorithm shows a relative robustness in these conditions. In contrast, the proposed TS-BASE/WF algorithm yields considerable SNR improvements at the left ear, and small SNR improvements at the right ear.

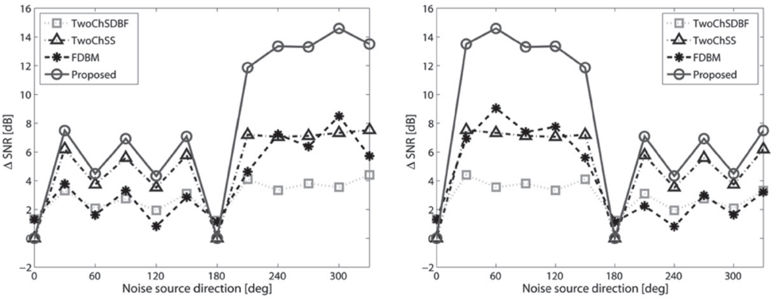

Fig. 2: SNR improvements (ΔSNRs) at the left ear (left) and the right ear (right) in one-noise-source conditions. TwoChSDBF: two-channel superdirective beamformer, TwoChSS: two-channel spectral subtraction, FDBM: frequency-domain binaural model, Proposed: two-stage binaural speech enhancement with Wiener filter (TS-BASE/WF).

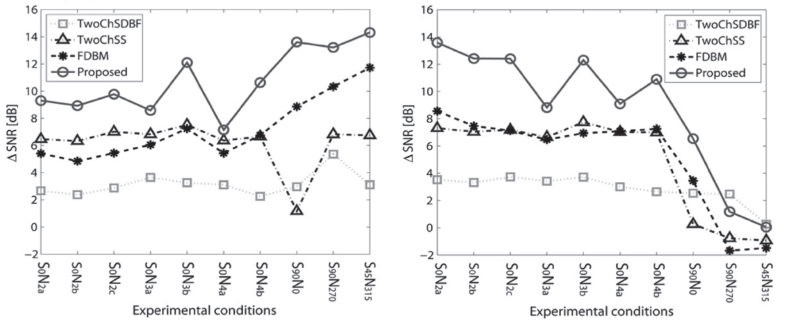

Fig. 3: SNR improvements (ΔSNRs) at the left ear (left) and the right ear (right) in multiple-noise-source conditions, and conditions with non-zero incoming direction of the target signal.

Objective evaluations for binaural cue preservation

For binaural processing, in addition to reducing interference components, the capability of preserving binaural cues is another important issue to evaluate. In objective evaluations for binaural cue preservation, the same target and interference signals as those used in the objective evaluations for speech enhancement were used. The noisy binaural signals were generated with a SNR of 0 dB under spatial configurations: the one-noise-source conditions ($S_{0:30:360}N_0$), and the three-noise-source conditions ($S_{0:30:360}N_{90,80,270}$), where the target source was simulated to be placed around the listener at positions from 0° to 360° in increments of 30°, and the interfering signal(s) were placed at fixed position(s).

Evaluation measure

The respective efficacies of the proposed TS-BASE/WF and other traditional algorithms in binaural cue preservation were evaluated objectively using the ITD error (E_{ITD}) and the ILD error (E_{ILD}) of the outputs.

The ITE error (E_{ITD}) is defined as

$$E_{\mathrm{ITD}} = \frac{\left|\angle c_{enhanced} - \angle c_{clean}\right|}{\pi} \quad , \tag{Eq. 2}$$

where $\angle c_{enhanced}$ and $\angle c_{clean}$ are the phases of the cross spectra (i.e., the approximate ITD estimates) for the enhanced and clean signals. Similarly, the ILD error (E_{ILD}) is defined as

$$E_{\mathrm{ILD}} = \left|10\log_{10} P_{enhanced} - 10\log_{10} P_{clean}\right| \quad , \tag{Eq. 3}$$

where $P_{enhanced}$ and P_{clean} respectively represent the power ratios (i.e., the approximate ILD estimates) for the enhanced signals and the clean signals.

Evaluation results

The results in E_{ITD} and E_{ILD} averaged across all utterances under the one-noise-source and three-noise-source conditions are shown respectively in Fig. 4 and Fig. 5. From Fig. 4, symmetry of EITD along with the median plane in the one-noise-source conditions is observed. Two facts contribute to this symmetric property: (1) symmetry of the HRIRs against the median plane; (2) operations in the spectral amplitude/ power domain of the studied algorithms. Regarding the comparisons of the studied algorithms, Fig. 4 illustrates that all studied algorithms exhibit different degrees of E_{ITD} under one-noise-source conditions. The traditional TwoChSS algorithm yields largest E_{ITD} after processing, which results from independent processing in two channels. The other traditional algorithms (i.e., TwoChSDBF and FDBM) introduce smaller E_{ITD} for the target signals with different arrival directions. These benefits are provided by the shared use of one filter with a real-value gain function at the left and right ears. The proposed TS-BASE/WF approach shows the smallest E_{ITD} under all tested spatial configurations. This virtue of the TS-BASE/WF algorithm can be attributed to: (1) the shared use of one filter in two channels; (2) its high noise reduction performance. The first factor enables preservation of the ITD cues of the binaural noisy input signals, and the second one significantly decreases the effects of interference components on the preserved ITD cues. The results in the three-noise-source conditions show that the traditional algorithms (TwoChSS, TwoChSDBF, and FDBM) again provide large E_{ITD}. Among the tested algorithms, the proposed TS-BASE/WF provides the smallest E_{ITD} in all tested conditions.

The results in E_{ILD} under one-noise-source and three-noise-source conditions are shown in Fig 5. Based on these results, it is observed that the TwoChSS algorithm shows the largest E_{ILD} in both one-noise-source and three-noise-source conditions because of the separate processing of binaural input signals. The traditional TwoChSDBF and FDBM algorithms demonstrate still high E_{ILD} in these conditions. The proposed TS-BASE/WF approach markedly reduces the ILD errors (i.e. the lowest E_{ILD}) due to the shared use of one filter in two channels and the high noise reduction capability.

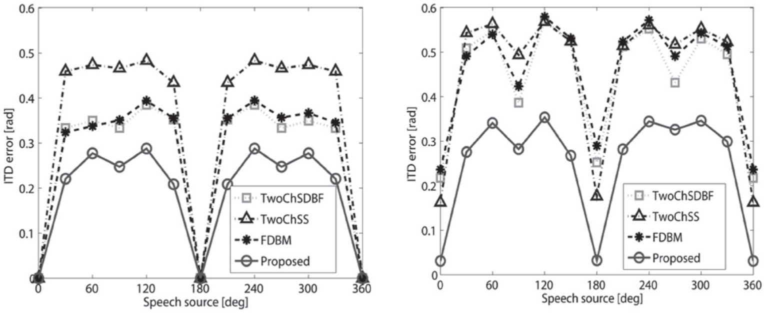

Fig. 4 The ITD errors in one-noise-source conditions ($S_{0:30:360}N_0$) (left) and three-noise-source conditions $S_{0:30:360}N_{90,180,270}$ (right).

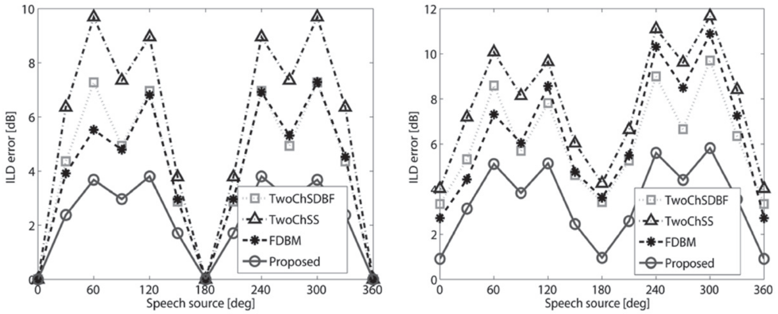

Fig. 5 The ILD errors in one-noise-source conditions ($S_{0:30:360}N_0$) (left) and three-noise-source conditions $S_{0:30:360}N_{90,180,270}$ (right).

CONCLUSIONS

In this paper, we objectively evaluated the two-stage binaural speech enhancement with Wiener filter (TS-BASE/WF) algorithm that we previously proposed through a number of experiments under different spatial configurations. The experimental results show that the TS-BASE/WF algorithm has two advantages: (1) effectiveness in dealing with non-stationary multiple-source interference signals, and (2) success in preserving binaural cues after processing. This proposed TS-BASE/WF can be potentially applied to realistic speech communication, hearing aid and so on.

ACKNOWLEDGEMENT

This research is partially supported by the Grant-in-Aid for Young Scientists (B) (No. 19700156) from the Ministry of Education, Science, Sports and Culture of Japan, and the SCOPE (071705001) of Ministry of Internal Affairs and Communication (MIC), Japan.

REFERENCES

Aichner R., Buchner H., Zourub M., and Kellermann W. (**2007**). "Multi-channel source separation preserving spatial information," Proc. ICASSP, pp. I.5-8.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785-797.

Dorbecker M., and Ernst S. (**1996**). "Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation," Proc. EUSIPCO, pp. 995-998.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking level differences," J. Acoust. Soc. Am. **35**, 1206-1218.

Jeffress, L. A. (**1948**). "A place theory of sound localization," J. Comparative and Physiological Psychology **41**, 35-39.

Klasen T. J., Van den Boqaert, T., Moonen, M., Wouters, J. (**2007**). “Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues processing,” IEEE Trans. on Signal Processing, **55**, 1579-1585.

Kollmeier B., Peissig J., and Hohmann V. (**1993**). “Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain,” Scand. Audiol. Suppl. **38**, 28-38.

Li, J., Sakamoto, S., Hongo, S., Akagi M., Suzuki Y. (**2009**). “Two-stage binaural speech enhancement with Wiener filter based on equalization-cancellation model,” Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics (New Paltz, NY, USA), pp. 133-136.

Lotter T., Sauert B., and Vary P. (**2005**). “A stereo input-output superdirective beamformer for dual channel noise reduction,” Proc., Eurospeech, pp. 2285-2288.

Nakashima H., Chisaki Y., Usagawa T., and Ebata M. (**2003**). “Frequency domain binaural model based on interaural phase and level differences,” Acoust. Sci. and Tech. **24**, 172-178.

Scalart P., and Filho J. V. (**1996**) “Speech enhancement based on a priori signal to noise estimation,” Proc. ICASSP, vol. 2, pp. 629-632.