euronoise

Two-input two-output speech enhancement with binaural spatial information using a soft decision mask filter

Satoshi Hongo[a], Ai Sasaki[b], Shuichi Sakamoto[b], Junfeng Li[c] and Yôiti Suzuki[b]

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

A two-input two-output speech enhancement method that preserves binaural spatial information in the output is preferred for realizing a comfortable auditory communication system. Such a system benefits from the noise reduction capability provided by signal processing technology, in addition to the binaural processing of the human auditory system.

We investigated a two-input two-output speech enhancement method that calculates soft decision mask filters to attenuate a noisy time-frequency bin. The soft decision mask filter is estimated for each direction of arrival (DOA) based on a noise-target ratio calculated using an adaptive filter that cancels the target signal. Results of computer simulations show that the proposed method has superior capabilities for maintaining spatial information in the two-output signals and for segregating the target signal in arbitrary azimuth and elevation DOA.

1 Introduction

People with normal hearing can enhance a target sound in a noisy environment up to 6–10 dB through binaural signal processing in an auditory system [1]. This is called selective binaural listening, and plays an important role in the cocktail party effect [1,2]. However, hearing impaired people have difficulty hearing in such conditions because of the degradation not only of the hearing threshold, frequency resolution, and temporal resolution, but also the acuity of selective binaural listening. Therefore, a speech enhancement algorithm for use in noisy environments is an indispensable component for hearing aids; the relevant literature includes many studies examining this problem [3-5].

On the other hand, even if binaural listening acuity is impaired, provision of proper binaural cues is important for hearing-impaired people because spatial perception is important to perceive the real world in daily life. Furthermore, the selective binaural listening acuity that hearing impaired people have, even if somewhat impaired, would improve the subjective signal to noise ratio (SNR), resulting in better speech intelligibility. In this context, as an assistive technology for impaired binaural hearing, two-output beamforming algorithms have been studied [6].

Not simple beamforming but speech enhancement signal processing with two-outputs seems effective to further actively cope with this issue. Moreover, considering applications for hearing aids, systems with two inputs must be the most feasible because two microphones can be placed near the listener's two ears. Therefore, two-input two-output speech enhancement algorithms seem promising for hearing aid applications because such systems can increase the noise reduction ability provided both by signal processing technology and by the selective binaural hearing processing of the human auditory system with reasonable system size [7].

Roman's two-input one-output system using a binary mask filter can be a good starting point [3] to develop such a two-input two-output speech enhancement algorithm. For the present study, we extended a binary mask to match a two-output system for binaural listening. The proposed system has two stages; the first stage applies an adaptive filter to estimate interference; the second stage applies a soft decision mask, i.e., a mask with decimal value of 0–1 to control the outputs precisely. The proposed method is called Two-Stage BinAural Speech Enhancement with a Soft Decision Mask Filter (TS-BASE/SDMF).

2 Two-input two-output speech enhancement algorithm

Figure 1 presents the concept of two-input two-output speech enhancement. The target speech signals from two input microphones near ears are enhanced with suppression of unnecessary noise. Two-output signals provide rich binaural cues for listeners to give full play of selective hearing and maintain proper spatial perception.

Figure 2 portrays a block diagram of the proposed system (TS-BASE/SDMF). The TS-BASE/SDMF has two stages: target cancellation through an adaptive filter (stage 1) and calculation of a soft decision mask filter (stage 2). It is assumed that the target signal comes from a certain known direction and that the interfering signals come from unknown directions. Moreover, no restrictions are imposed on the number, location, or contents of the interference.

In stage 1, the interference components are estimated through an adaptive filter that cancels the target signal. The active filter should fulfill the law of causality to cancel the target signal properly. Therefore, according to the direction of arrival (DOA) of the target sound source, the channel to insert an active filter was altered electronically, as depicted in Fig. 3. For example, if the right side source is to be enhanced, the right signal is input earlier than the left corresponded signal. Therefore, the active filter must be switched to the right side.

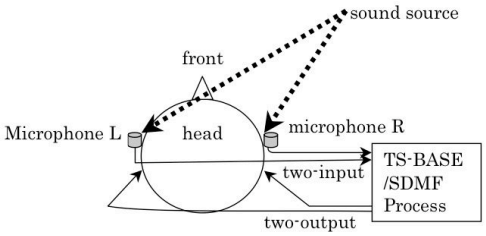


Fig. 1 Locations of head and microphone

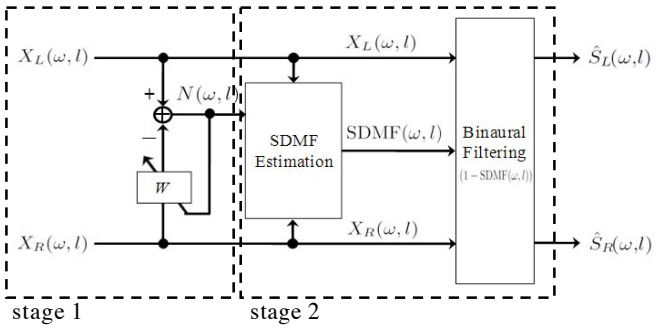
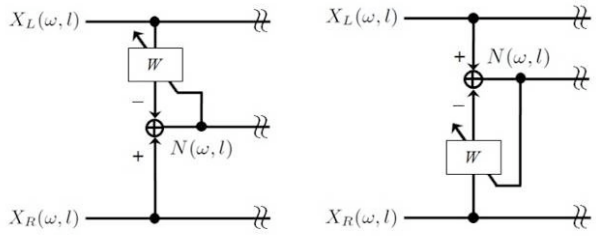


Fig. 2 Block diagram of TS-BASE/SDMF



(a) Target is the left side (b) Target is the right side

Fig. 3 Switching the Active filter *w* in stage 1

In stage 2, a soft decision mask filter ranging continuously from 0 to 1 is given based on the estimated noise-target ratio $R(\omega,l)$ that is defined as

$$R(\omega,l)=\frac{|N(\omega,l)|}{\left(|X_L(\omega,l)|+|X_R(\omega,l)|\right)/2}\ . \qquad (1)$$

Here, $N(\omega,l)$ is the estimated interference component through the adaptive filter. In addition, $X_L(\omega,l)$ and $X_R(\omega,l)$ respectively represent the observed microphone signals at the listener's left and right ears.

Using Eq. (1), when the input signal is dominated by interference, then $R(\omega,l)\approx 1$. On the other hand, when the input signal contains the target signal, then $R(\omega,l)<<1$. A soft decision mask filter (SDMF) can be implemented based on $R(\omega,l)$ as the following.

$$SDMF(\omega,l)=\begin{cases}1 & R(\omega,l)>R_{\max}(X_L,X_R,N),\\ 0 & R(\omega,l)<R_{\min}(X_L,X_R,N),\\ \dfrac{R(\omega,l)-R_{\min}}{R_{\max}-R_{\min}} & R_{\min}<R(\omega,l)<R_{\max}.\end{cases} \qquad (2)$$

The value of $R_{\max}$ and $R_{\min}$ can be changed based on the relation of $X_L(\omega,l)$, $X_R(\omega,l)$ and $N(\omega,l)$. Finally, the enhanced signals are calculated as

$$\hat{S}_i(\omega,l)=\left(1-SDMF(\omega,l)\right)\cdot|X_i(\omega,l)|\cdot\exp(j\angle X_i(\omega,l)),$$
where $i=\mathrm{L\ or\ R}$. (3)

In Roman's algorithm, the denominator in Eq. (1) is the input signal at the primary microphone (the ear with a higher SNR). Therefore, the performance degrades when the positions of the interference change. Furthermore, they applied a hard binary mask taking a value of either 0 or 1. The performance, using a soft decision mask filter with any decimal value from 0 to 1, would be better than that using a binary mask because the speech sounds processed with the binary mask usually generate musical noise and the output sound quality often becomes unnatural.

Comparing TS-BASE/SDMF to Roman's method, advantageous characteristics of TS-BASE/SDMF might be summarized as follows: (1) perceptual gain provided by human selective binaural hearing and proper spatial perception based on two channel outputs, (2) better sound quality offered by soft decision mask filters which can control the outputs moderately in the time (frame) and frequency domain, and (3) automatic decision of the reference microphone using the average of two input signals.

3 Performance evaluation by computer simulation

3.1 Condition of Simulations

The frame length and the frame overlap were set respectively to 512 and 128 taps, where the sampling frequency was 16 kHz. A Hanning window was used for the framing. The normalized least mean square (NLMS) algorithm was used for adaptive filtering. White noise was used to calibrate the filter in the absence of interference. After the training phase of 20 s, the filter coefficients were fixed and applied to the observed input signal in the presence of interference.

The locations of the sound sources in our computer simulations were simulated by convolving the Head-Related Impulse Responses (HRIRs) of a KEMAR Dummy Head [8]. The distance between the loudspeaker and receiver was 1.4 m. The elevation of the sound direction was set to 0°. The SNR was set to -5 dB to 10 dB by 5 dB steps. The input SNR was calculated at the left ear after addition of all interference signals. Figure 4 shows the DOAs of sound sources that were used for the computer simulation. The target signal was a male utterance issuing from 0° (frontal incidence). As portrayed in Fig. 4, the tested conditions were as follows: (1) interference from a female speaker at 45° (Condition 1); (2) four concurrent speakers (two female and two male utterances) at azimuth angles of -135°, -45°, 45° and 135° (Condition 2).

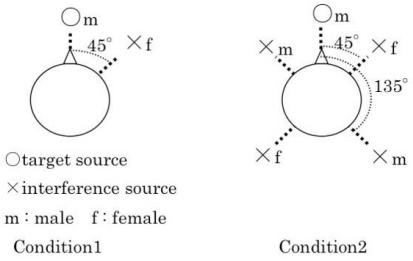


Fig. 4 Sound source location for simulation

3.2 SNR and LSD evaluation

We selected the objective quality measure known as Segmental SNR [9] and log-spectral distance (LSD) [10] to evaluate TS-BASE/SDMF. The higher the Segmental SNR is, the higher the speech quality is. The lower the LSD is, the lower the speech distortion is. Segmental SNR and LSD were calculated as

$$\mathrm{Seg.SNR}=\frac{10}{L}\sum_{l=0}^{L-1}\log_{10}\left(\frac{\sum_{k=0}^{K-1}[s(lK+k)]^2}{\sum_{k=0}^{K-1}[\hat{s}(lK+k)-s(lK+k)]^2}\right), \qquad (4)$$

$$\mathrm{LSD}=\frac{10}{L}\sum_{l=0}^{L-1}\left(\frac{1}{K}\sum_{k=0}^{K-1}[\log_{10}AS(k,l)-\log_{10}A\hat{S}(k,l)]^2\right)^{1/2}, \qquad (5)$$

where $AS(k,l)=\max\left\{|s(k,l)|^2,\delta\right\}$, (6)

and

$$\delta=10^{-\frac{50}{10}}\max\left\{|S(k,l)|^2\right\}. \qquad (7)$$

Moreover, s and S respectively signify the target signal waveform and spectrum, and $\hat{s}$ and $\hat{S}$ respectively denote those of the segregated signal; L and K respectively represent the numbers of frames and taps.

Results of simulations of several SNR conditions are presented in Table 1 and Table 2. Values of $R_{\max}$ and $R_{\min}$ were set respectively to 0.7 and 0.3. The results show that the segmental SNR of TS-BASE/SDMF is higher than that of Roman's algorithm by about 1–3 dB. The LSD of TS-BASE/SDMF is around the same as that of Roman's.

Table 1 Results for one-speaker interference (Condition 1)

Input SNR		-5 dB	0 dB	5 dB	10 dB
Input SegSNR (dB)		-5.6	-0.7	4.3	9.2
SegSNR (dB)	Roman's algorithm	4.6	6.2	8.0	9.9
	TSBASE /SDMF	6.5	8.4	10.	13.
Input LSD (dB)		7.8	7.0	5.1	3.6
LSD(dB)	Roman's algorithm	4.9	4.1	3.3	2.7
	TSBASE /SDMF	4.8	4.0	3.2	2.6

Table 2 Results for four-speaker interference (Condition 2)

Input SNR		-5 dB	0 dB	5 dB	10 dB
Input SegSNR(dB)		-12.	-6.5	-1.5	3.5
SegSNR (dB)	Roman's algorithm	-2.7	0.4	3.4	6.5
	TSBASE /SDMF	-1.7	1.4	4.5	7.8
Input LSD (dB)		11.	8.7	6.3	4.4
LSD(dB)	Roman's algorithm	6.1	5.0	3.9	2.9
	TSBASE /SDMF	5.9	4.8	3.6	2.8

3.3 Influence of sound sources and acoustical conditions

A simulation in which the direction of the interference (white noise) is changed from 0° to 360° was performed to investigate the influence of the location of interference signal (noise) (Fig. 5). The elevation of the sound direction was set to 0°, with the target male speaker at 0° (Figs. 6(a) and 6(b)) and 60° (Figs. 6(c) and 6(d)). The performance was evaluated at the left output channel; the SNR at the sound source is set to -5 dB.

Figure 6 shows results of Segmental SNR and LSD. In both Roman's method and TS-BASE/SDMF, the transfer function difference of left and right channels was used to determine the mask filter. Therefore, if the position of target and noise are the same, the adaptive filter to suppress the target signal will become the filter to suppress interference. As a result, the noise-target ratio $R(\omega,l)$ cannot be calculated correctly (0° and 60°). In fact, when DOAs of interference and target are close, the performance of TS-BASE/SDMF generally degrades, as observed from data from 30° to 150° (left side) in Fig. 6. In addition, the symmetrical position about the two-input sensor also becomes the position with low performance (180° and 120°).

In fact, TS-BASE/SDMF outperforms Roman's method if the target source position is set to the opposite side of the noise position. Particularly when the target is 60°, the performance difference is observed clearly.

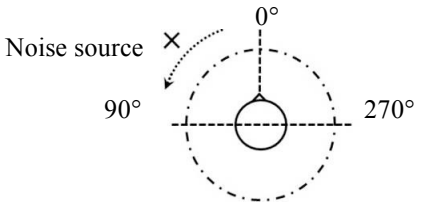
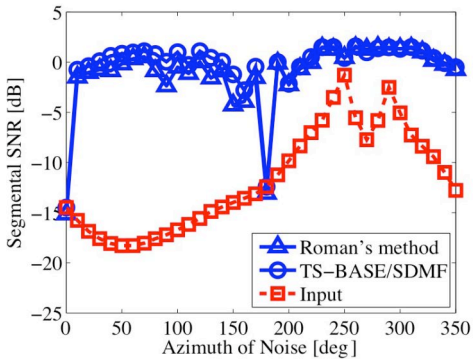
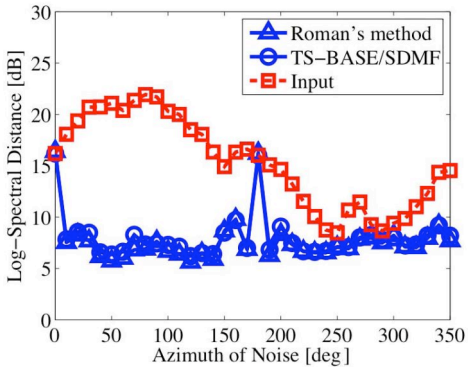


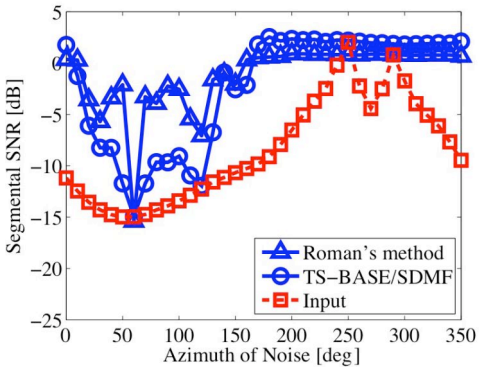
Fig. 5 noise rotation around the head



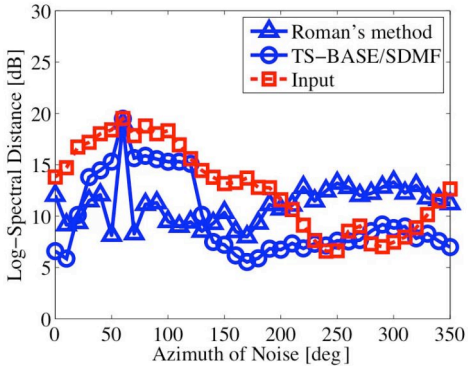
(a) Segmental SNR (target = 0°)



(b) LSD (target = 0°)



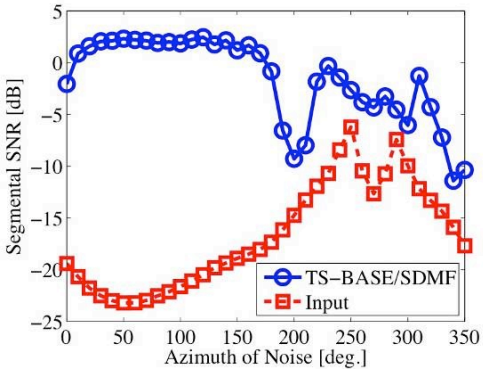
(c) Segmental SNR (target = 60°)



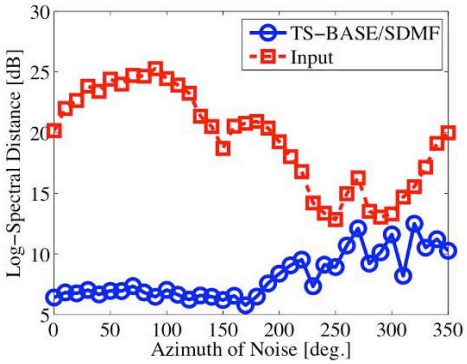
(d) LSD (target = 60°)

Fig. 6 Segmental SNR and LSD when the noise signal direction was changed.

Figure 7 shows results of the simulation when the elevation and azimuth were set respectively to 60° and 96°. Results clarified that the proposed method can enhance the target signal for all azimuths, even if the elevation of target sound DOA is other than zero.



(a) Segmental SNR



(b) LSD

Fig. 7 Simulation results showing that the elevation of DOA $\neq 0$ (elevation and azimuth were set to 60° and 96°)

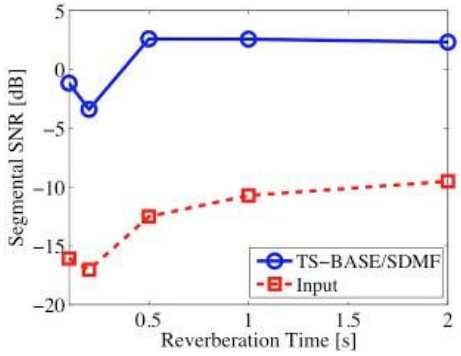
The influence of the performance caused by reflections must be evaluated to use TS-BASE/SDMF in reverberant rooms. A simulation with was carried out with various reverberant times. The sound direction elevation is set to 0°; the azimuth angle of the target and interference were set to 0° and 60°. Reverberation was added by convolving exponentially decaying random noise.

Figure 8 depicts the results of Segmental SNR and LSD in a reverberant environment. In fact, TS-BASE/SDMF exhibits much better characteristics throughout the examined reverberant time range, which might mean that TS-BASE/SDMF is useful for practical applications in rooms with moderate reverberation time.

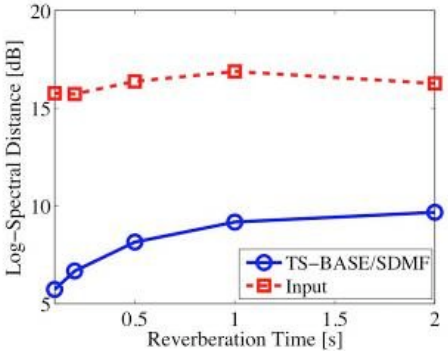
3.4 Evaluation for maintaining spatial information

TS-BASE/SDMF uses the same two segregation filters between left and right channels. Therefore, interaural phase difference (IPD) and interaural level difference (ILD) of the target source must be well preserved in the case of high SNR. It is necessary to confirm how well IPD and ILD are preserved in practical environments with low SNR. The elevation and azimuth angle of the target signal are 0° and 300°, respectively. The azimuth angle of the interference was set at 60°. The SNR before the sound source and interference were convoluted with HRIR was set to 0 dB.

Figure 9 shows that the result that IPD and ILD of segregated speech are nearly equal to those of the target speech, even at SNR of 0 dB.

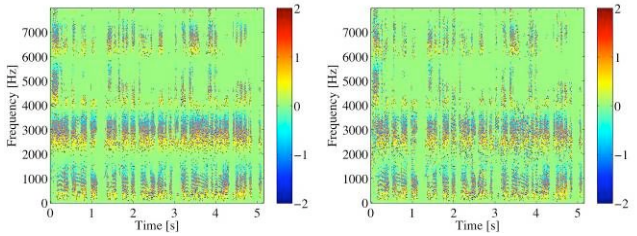


(a) Segmental SNR

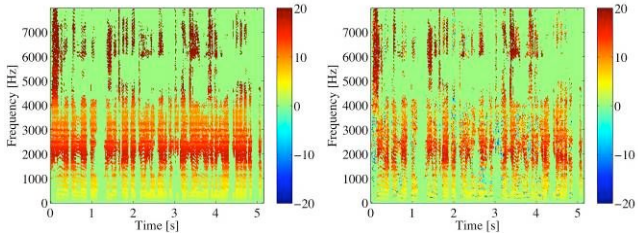


(b) LSD

Fig. 8 Performance in reverberant environment
Azimuths of target and interference were set to 0° and 60°



(a) IPD of target speech [π rad] (b) IPD of segregated speech

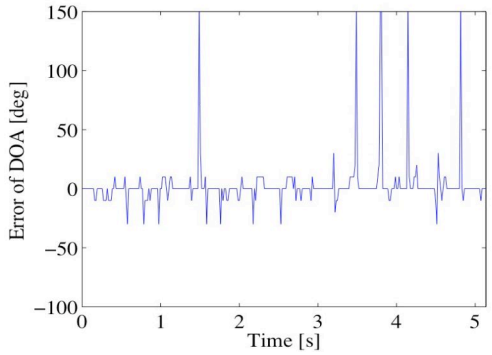


(a) ILD of target speech [dB] (b) ILD of segregated speech

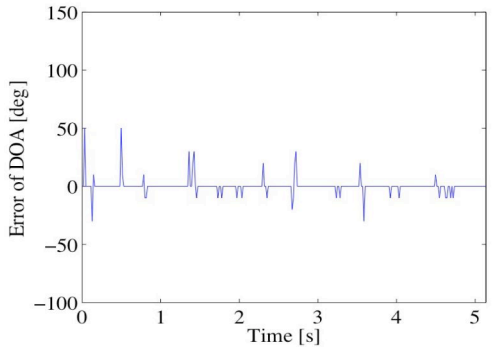
Fig. 9 IPD and ILD of target speech and segregated speech

The DOA estimation of the enhanced signal was compared to the real direction of the target signal to evaluate performance quantitatively. DOA was calculated using a procedure like that of [4]: (1) the relation map between DOA and IPD (or ILD) as a function of direction was built based on a head-related transfer function (HRTF); (2) DOA was estimated by comparing IPD (or ILD) calculated from the segregated signal to those of the relation map; (3) the DOA error was given as the difference between the estimated DOA of target and real target direction. The DOA was estimated using IPD for low frequencies of less than 750 Hz, and using ILD at high frequencies greater than 1500 Hz.

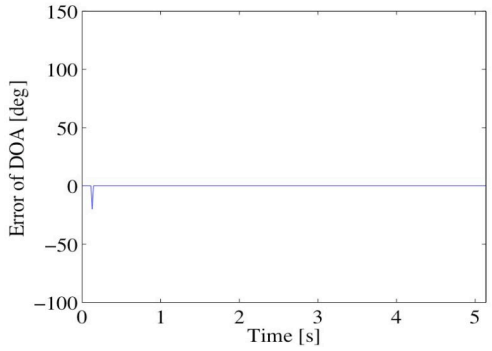
Figure 10 shows the time course of DOA error at various frequencies. At low frequencies, large and pulsive DOA errors are observed. The pulsive errors seem attributable to the timings where the target speech were started and ended. These DOA errors are probably caused by phase control errors in speech frame-frequency bins when switching the frames. Because the DOA error becomes almost zero when the speech signal is steady, the proposed method has capabilities for maintaining spatial information in the two-output signals.



(a) 500 Hz



(b) 2000 Hz



(c) 5000 Hz

Fig. 10 DOA error

(target 300°, interference 60°, SNR 0 dB)

5 Conclusion

We proposed a new two-input two-output speech enhancement algorithm called TS-BASE/SDMF, which includes the two stages of (1) target cancellation through adaptive filtering, and (2) speech enhancement through a soft decision mask filter (SDMF).

The TS-BASE/SDMF algorithm has theoretically advantageous characteristics: (1) good sound quality by SDMF in place of a hard binary mask; (2) rich and proper binaural cues provided by two outputs.

Results of computer simulations showed that TS-BASE/SDMF shows good performance in the following points: (1) enhancement of the target signal in wide azimuth and non-zero elevation of target sound DOA; (2) robustness for reverberation in acoustical environments; (3) good capabilities for maintaining spatial information (binaural cues) in the two-output signals.

Subjective evaluation of how well binaural cues are preserved in TS-BASE/SDMF is an interesting task for future study.

Acknowledgment

This research was supported in part by the Global COE Program Center of Education and Research for Information Electronics Systems (CERIES).

References

[1] Masanao Ebata, “Spatial unmasking and attention related to the cocktail party problem,” Acoust. Sci. and Tech., Vol. 24, No. 5, pp. 208–219, 2003.

[2] B. C. J. Moore, “An Introduction to the Psychology of Hearing (3rd Edition)” (1989).

[3] Nicoleta Roman, Soundarara jan Srinivasan, and DeLiang Wang, “Binaural segregation in multi source reverberant environments,” J. Acoust. Soc. Am., Vol. 120, No. 6, pp. 4040–4051, 2006.

[4] Hidetoshi Nakashima, Yoshifumi Chisaki, Tsuyoshi Usagawa, and Masanao Ebata, “Frequency domain binaural model based on interaural phase and level differences,” Acoust. Sci. and Tech., Vol. 24, No. 4, pp. 172–178, 2003.

[5] Junfeng Li, Shuichi Sakamoto, Satoshi Hongo, Masato Akagi, and Y^oiti Suzuki, “A Speech Enhancement Approach for Binaural Hearing Aids,” Proc. of 22nd SIP Symposium (2007) 263.

[6] Y\^oiti Suzuki, Shinji Tsukui, Futoshi Asano, Ryouichi Nishimura, and Toshio Sone, “New design method of a binaural microphone array using multiple constraints,” IEICE TRANS. on Fundamentals of Electronics, Communications and Computer Sciences, E82-A (4) (1999), 588-596.

[7] Ryouichi Nishimura, Y\^oiti Suzuki, Shinji Tsukui, Futoshi Asano, “Array signal processing with two outputs preserving binaural information,” Applied acoustics, 65 (7) (2004), 657-672.

[8] Bill Gardner and Keith Martin “HRTF Measurements of a KEMAR Dummy-Head Microphone,” URL: http://sound.media.mit.edu/KEMAR.html, 1994http://sound.media.mit.edu/KEMAR.html.

[9] Junfeng Li and Masato Akagi, “Noise reduction method based on generalized subtractive beamformer,” Acoust. Sci. and Tech., Vol. 27, No. 4, pp. 206–215, 2006.

[10] Israel Cohen, “Multi channel Post-Filtering in Non-stationary Noise Environments,” IEEE Trans. Signal Processing, Vol. 52, No. 5, pp. 1149–1160, 2004.