Neurocomputing 72 (2008) 636–642

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Letters

# Binaural semi-blind dereverberation of noisy convoluted speech signals

Jong-Hwan Lee [a,*,1], Sang-Hoon Oh [b], Soo-Young Lee [a]

[a] Brain Science Research Center and Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[b] Department of Information Communication Engineering, Mokwon University, Daejeon, Republic of Korea

ARTICLE INFO

*Article history:*
Received 17 October 2007
Received in revised form
3 July 2008
Accepted 6 July 2008
Communicated by T. Heskes
Available online 9 September 2008

*Keywords:*
Independent component analysis
Blind dereverberation
Blind deconvolution
Blind least squares
Speech enhancement
Automatic speech recognition

ABSTRACT

In order to overcome a limited performance of a conventional monaural model, this letter proposes a binaural blind dereverberation model. Its learning rule is derived using a blind least-squares measure by exploiting higher-order characteristics of output components. In order to prevent an unwanted whitening of speech signal, we adopt a semi-blind approach by employing a pre-determined whitening filter. The proposed model is evaluated using several simulated conditions and the results show better speech quality than those of the monaural model. The applicability of the model to the real environment is also shown by applying to real-recorded data. Especially, the proposed model attains much improved word error rates from $13.9 \pm 5.7(\%)$ to $4.1 \pm 3.5(\%)$ across 13 speakers for testing in the real speech recognition experiments.

© 2008 Elsevier B.V. All rights reserved.

## 1. Introduction

Speech signals convolved with room reverberations become corrupted and the performance of an automatic speech recognition (ASR) system is subsequently degraded [3]. In order to restore clean speech signal, blind dereverberation (BD) methods have been proposed by estimating an inverse or dereverberation filter of the unknown room reverberations from a recorded signal [9,4–6,8,13].

Fig. 1a shows a conventional single-channel (monaural) model based on an assumption that a speech signal S has a non-Gaussian probability density function (PDF). In order to prevent decorrelation among speech samples (i.e. whitening), a reverberated signal X is pre-processed with a pre-trained whitening filter $W_t$ and a whitened signal $X_t$ is used to estimate a dereverberation filter W (i.e. semi-blind dereverberation or sBD). By exploiting higher-order characteristics of a dereverberated signal $U_t$, the dereverberation filter W is updated based on a gradient term defined as [9,4,8]

$$\Delta W = \left( \frac{1}{W^*} - STFT(\varphi(u_t)) \bullet X_t^* \right) \bullet W^* \bullet W, \tag{1}$$

where a function *STFT* denotes a short-time fast Fourier transform (STFT), $\bullet$ is an element-wise multiplication between two vectors, $^*$ is a complex conjugate operation, and $\varphi(u_t) \triangleq -\partial \ln p(u_t)/\partial u_t$ is a minus of the Fisher score function [7] ($p(u_t)$: a PDF of $u_t$). Then, the dereverberation filter is iteratively updated as follows:

$$W_{new} = W + \eta \Delta W, \tag{2}$$

where a positive scalar $\eta$ ($\ll 1$) is a learning rate.

However, due to non-minimum phase characteristics of room reverberation, an exact inverse filter cannot be achieved from the single-channel model and an additional channel is required [6,17]. In this context, this letter proposes a dual-channel (binaural) sBD model shown in Fig. 1b. In a reverberation block, a clean speech signal S is convolved with two reverberation channels $\mathbf{H}_i$ ($i \in \{1,2\}$). The resulting convolved signals are further corrupted with additive white Gaussian noises (AWGN) $\mathbf{N}_i$ that model measurement noises during recording processes. In a concurrent dereverberation block, we may assume that there are two independent sources (i.e. S and $\mathbf{N}_i$) and, subsequently, the binaural sBD model can be adopted from a blind source separation (BSS) model. Using a blind least-squares (BLS) as a measure of independence among output components [9,2,1], an iterative learning rule of the corresponding model is derived by minimizing the BLS and thus by maximizing independence across and within output components. Upon evaluation using simulated reverberations, the proposed binaural sBD model is applied to a real-recorded data.

* Corresponding author.
*E-mail address:* jhlee.jonghwanlee@gmail.com (J.-H. Lee).
[1] Now at Brigham and Women's Hospital, Harvard Medical School, 75 Francis St, Boston, MA 02115, USA.

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.07.005

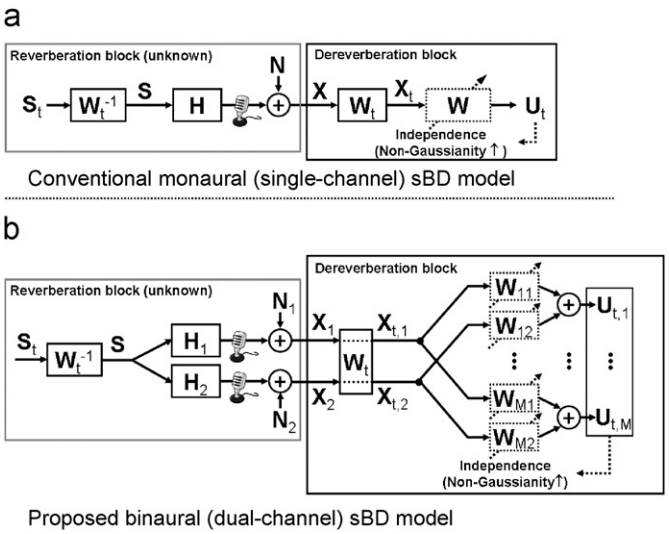



**Fig. 1.** Block diagrams of semi-blind dereverberation (sBD) models: (a) a conventional monaural (single-channel) and (b) proposed binaural (dual-channel) models.

## 2. A learning algorithm of the proposed binaural model

Using a STFT, the whitened reverberated signal $X_{t,i}$ in Fig. 1b is expressed as

$$X_{t,i} = W_t \bullet X_i = W_t \bullet (H_i \bullet W_t^{-1} \bullet S_t + N_i) = H_i \bullet S_t + W_t \bullet N_i, \quad (3)$$

where $i$ ($\in \{1,2\}$) is an index of microphone. Therefore, $H_i$ is the only convolution channel in $X_{t,i}$ and $W_t \bullet N_i$ is a filtered Gaussian noise. By applying dereverberation filters,

$$U_{t,j} = \sum_{i=1}^{2} (W_{ji} \bullet X_{t,i}), \quad (4)$$

where $j$ ($\in \{1,\ldots,M\}$) is an index of output and $W_{ji}$ is a STFT of an $L$th-order finite impulse response (FIR) filter $w_{ji}$ from the $i$th microphone to the $j$th output component. Because of an unwanted problem of circular convolution, the first $L$ samples of $u_{t,j}$ ($= STFT^{-1}(U_{t,j})$) should be discarded.

As a measure of independence among output components, we adopt a BLS cost function defined as [9,2,1]

$$J = \sum_{j=1}^{M} E[|U_{t,j} - STFT(g(u_{t,j}))|^2] \quad (5)$$

in the frequency domain, where $g(\cdot)$ is a nonlinear function related with a PDF of source signal [1,14]. In order to minimize Eq. (5) based on a stochastic gradient descent method [16], an iterative learning rule can be derived by taking the derivative of Eq. (5) with respect to $W \triangleq [W_{ji}]_{M\times 2}$. Additionally, a natural gradient scheme is applied to improve convergence property [4,8] and an error term ($u_{t,j} - g(u_{t,j})$) is substituted as $\varphi(u_{t,j})$ for asymptotical efficiency [1,14]. Finally, a learning algorithm for the proposed binaural sBD model can be derived as

$$\Delta W = -\frac{\partial J}{\partial W^*} \bullet W^H \bullet W \approx -\begin{bmatrix} STFT(\varphi(u_{t,1})) \\ \vdots \\ STFT(\varphi(u_{t,M})) \end{bmatrix} \bullet \begin{bmatrix} U_{t,1} \\ \vdots \\ U_{t,M} \end{bmatrix}^H \bullet W, \quad (6)$$

where H is a Hermitian transpose.

To avoid a trivial zero solution of Eq. (6), W is normalized to unit norm on every iteration. Therefore, a newly updated term $W_{new}$ can be represented as

$$W_{new} = \frac{W + \eta \Delta W}{\|W + \eta \Delta W\|} \triangleq W + \Delta W', \quad (7)$$

where $\|\cdot\|$ denotes an $L_2$-norm and $\eta$ is a learning rate ($\eta \ll 1$). By substituting Eq. (6) into $\Delta W$ of Eq. (7),

$$\Delta W' = \frac{\eta}{1-\mu\eta} \left\{ \mu \begin{bmatrix} \underline{\mathbf{1}} & & \underline{\mathbf{0}} \\ & \ddots & \\ \underline{\mathbf{0}} & & \underline{\mathbf{1}} \end{bmatrix}_{(M\times M)} - \begin{bmatrix} STFT(\varphi(u_{t,1})) \\ \vdots \\ STFT(\varphi(u_{t,M})) \end{bmatrix}_{(M\times 1)} \bullet \begin{bmatrix} U_{t,1} \\ \vdots \\ U_{t,M} \end{bmatrix}_{(M\times 1)}^H \right\} \bullet W_{(M\times 2)}, \quad (8)$$

where $\underline{\mathbf{0}}$ is a zero vector, $\underline{\mathbf{1}}$ is a vector whose elements are all 1's, $\mu \triangleq 1 - \|W + \eta\Delta W\|/\eta$, and a dimension of each matrix is noted as a subscript inside a parenthesis. Note that due to $2^N$-point STFT, each element in Eq. (8) is a $2^N$ dimensional vector. Since each element of $\Delta W'$ goes to $\underline{\mathbf{0}}$ at the convergence,

$$STFT(\varphi(u_{t,j})) \bullet U_{t,j}^* = \mu \underline{\mathbf{1}} \quad \text{and} \quad STFT(\varphi(u_{t,j})) \bullet U_{t,k}^* = \underline{\mathbf{0}} \ (j \neq k), \quad (9)$$

where $j,k \in \{1,\ldots,M\}$.

According to a higher-order decorrelation property of Eq. (9), each output would be dereverberated based on the first constraint (i.e. statistically independent sequence) and the resulting $M$ outputs would be independent of each other based on the second constraint. Therefore, we can anticipate that AWGN may be extracted by one of the outputs and speech signals may be decomposed into the remaining outputs. Consequently, we tested the binaural model for two different output numbers ($M \in \{2,3\}$). Note that, due to spectral dependency of adjacent speech samples, the resulting speech components may be separated from frequency-dependent components [10].

An overall procedure for iterative training of the dereverberation filters $w_{ji}$ can be summarized as follows:

(i) Transform the time domain $L$th-order filter $w_{ji}$ and $K$ samples of $x_{t,i}$ into the frequency domain representations of $W_{ji}$ and $X_{t,i}$ using a $2^N$-point STFT, respectively. Here, $K > L$ and $N$ is the nearest integer such that $2^N \geqslant K$.
(ii) Calculate the output components of $U_{t,j} (= \sum_{i=1}^{2} (W_{ji} \bullet X_{t,i}))$ and $u_{t,j}$ ($= STFT^{-1}(U_{t,j})$). Note that the first $L$ samples of $u_{t,j}$ should be discarded due to an unwanted problem of a circular convolution.
(iii) Update $W_{ji}$ using Eq. (7) and obtain $w_{ji}$ by taking the first $(L+1)$ real values of $STFT^{-1}(W_{ji})$.
(iv) Iterate (i)–(iii) until a pre-defined stopping criterion is reached.

The proposed dereverberation algorithm of Eq. (6) was derived using the BLS cost function. As some of the machine learning problems, the adopted cost function may be non-convex over the region of operating parameters or may have multiple local minima [11]. Subsequently, it is complicated to find the exact optimal solution of the problem which can be derived from a convex cost function. Instead, without a complication of a global optimization process [12], the problem can be solved by applying adaptive approaches such as the adopted stochastic gradient descent method which finds a minimum point based on a negative gradient of the cost function [16]. Note that the convergence to the global minimum may not be guaranteed due to the existence of multiple local minima and even saddle points, and thus an additional work on the convergence analysis is warranted.

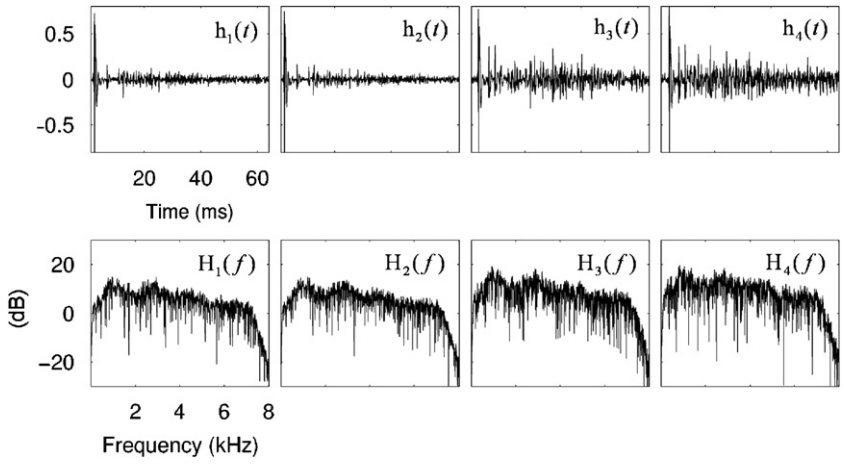



**Fig. 2.** Temporal waveforms and magnitude spectra of four employed simulated reverberation channels ($5 \times 4 \times 3\,m^3$ sized room; one speaker at (2.0 m, 2.0 m, 1.0 m); and two pairs of microphones at $m_1$:(1.9 m, 1.5 m, 1.0 m) & $m_2$:(2.1 m, 1.5 m, 1.0 m) and $m_3$:(2.0 m, 1.0 m, 1.0 m) & $m_4$:(2.5 m, 1.0 m, 1.0 m); $h_i$ corresponds to a reverberation channel measured at $m_i$).

## 3. Evaluation using simulated reverberations

In the TIMIT speech corpus,[2] data from a randomly selected speaker 'mjwt0' (~30 s; 16 kHz sampling) were used to train both whitening and dereverberation filters. That is a speaker dependent condition. A FIR whitening filter (1024-tap, 64 ms; 512-tap delay), pre-trained using the monaural algorithm presented in Eqs. (1) and (2), was applied to both monaural and binaural models. Four reverberations shown in Fig. 2 were taken using a commercial software 'Room Impulse Response v2.5'[3] which employs a time-domain image expansion method. A reverberated speech signal was further corrupted with AWGN of 20, 15, and 10 dB signal-to-noise ratio (SNR) levels.

In order to measure speech quality, perceptual evaluation of speech quality (PESQ) mean opinion score (MOS)[4] was employed (ranges: 1−5; 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent). In Fig. 1b 4096-tap (256 ms) FIR filters with 2048-tap delay were used as $w_{ji}$. As a semi-batch learning scheme, $w_{ji}$ was updated every 8192-sample of $u_{t,j}$ which was obtained from 12 287-sample ($= 4096 + 8192 - 1$) of $x_{t,i}$. A 16 384-point STFT was used, and in each sweep, $\eta$ was adaptively changed so that an averaged energy of ($\eta \Delta W/W$) was fixed at $10^{-4}$.

Fig. 3 shows total channel responses of considered models after the convergence of PESQ MOS (condition: $h_3$ and $h_4$ with 15 dB SNR). For the binaural model, the total channel response at the $j$th output was defined as

$$a_j(t) \triangleq h_3(t) * w_{j1}(t) + h_4(t) * w_{j2}(t) \quad \& \quad A_j(f) \triangleq STFT(a_j(t)). \tag{10}$$

Comparing $h_4$ in Fig. 2 with $a(t)$ in Fig. 3a, although a large amount of distortions from reverberation channel was removed by the monaural model, 'zeros' still remained in $A(f)$. From the results of the binaural model ($M = 2$) in Fig. 3b, speech signal and Gaussian noise were separated into the 1st and 2nd components, respectively. Note that from the $A_1(f)$ within 0−4 kHz, the remaining zeros of the monaural model were successfully removed. Similarly, as shown in Fig. 3c, the proposed model ($M = 3$) decomposed $x_{t,i}$ into noise component (the 2nd) and two speech components (the 1st and 3rd). Interestingly, we can see that $a_1(t)$ and $a_3(t)$ showed low- and high-pass filter characteristics, respectively (~4 kHz cut-off). This 'frequency division' occurred due to the strong temporal and spectral dependencies of adjacent speech samples as described in Section 2. Again, compared to the monaural model, the proposed model ($M = 3$) showed much improved total channel response within 0−4 kHz without spectral zeros (also slightly better than the results from $M = 2$).

For each sweep, PESQ MOS of a dereverberated speech was measured within 0−4 kHz because most of phonetic features for ASR are extracted within this frequency range [15]. Accordingly, the 16 kHz data were down-sampled to 8 kHz for the monaural model. Since the reverberation channels were also reduced from 1024- to 512-tap from the down-sampling, a 2048-tap FIR filter (1024-tap delay) was used as w for the monaural model. So, the ratio between reverberation and dereverberation filters was same in the monaural model (512-tap h vs. 2048-tap w) and binaural model (1024-tap h vs. 4096-tap $w_{ji}$).

Fig. 4 shows the learning curves of PESQ MOS values, which were averaged across reverberation channels. Note that the proposed binaural model with $M = 3$ showed better PESQ MOS values compared to the proposed binaural model with $M = 2$ as well as the conventional monaural model. This result is consistent with the total channel responses shown in Fig. 3. Additionally, a dominant improvement of the proposed model for a low SNR indicates that our model can also reduce additive noises in the dereverberated speech signals (i.e. compare monaural/binaural ($M = 2$) models for 20 and 10 dB SNRs in Fig. 4).

## 4. Results of real-recorded data

We tested the proposed binaural model with $M = 3$ on a real recording environment in an office room sized $4.5 \times 4.5 \times 2.5\,m^3$. A 75 Korean phonetically balanced isolated-word (PBW) database (~60 s per speaker) was played by a normal PC speaker and was recorded using two condenser microphones (ATR 35 s; Audio-technica) and a Sound Blaster PCI128 card (16 kHz sampling). A distance between the speaker and one of the microphones was 100 cm (distance between two microphones: 20 cm). An FIR whitening filter (1024-tap) was pre-trained using 35 speakers' clean speech signals. The speech data of the remaining 13 speakers were used for testing the proposed model (i.e. speaker independent condition).

The dereverberation filters were iteratively updated following the training process described in Section 2. FIR filters (4096-tap) were used as $w_{ji}$. We also tested FIR dereverberation filters that

[2] http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html
[3] http://www.dspalgorithms.com/room/room25.html
[4] http://www.pesq.org

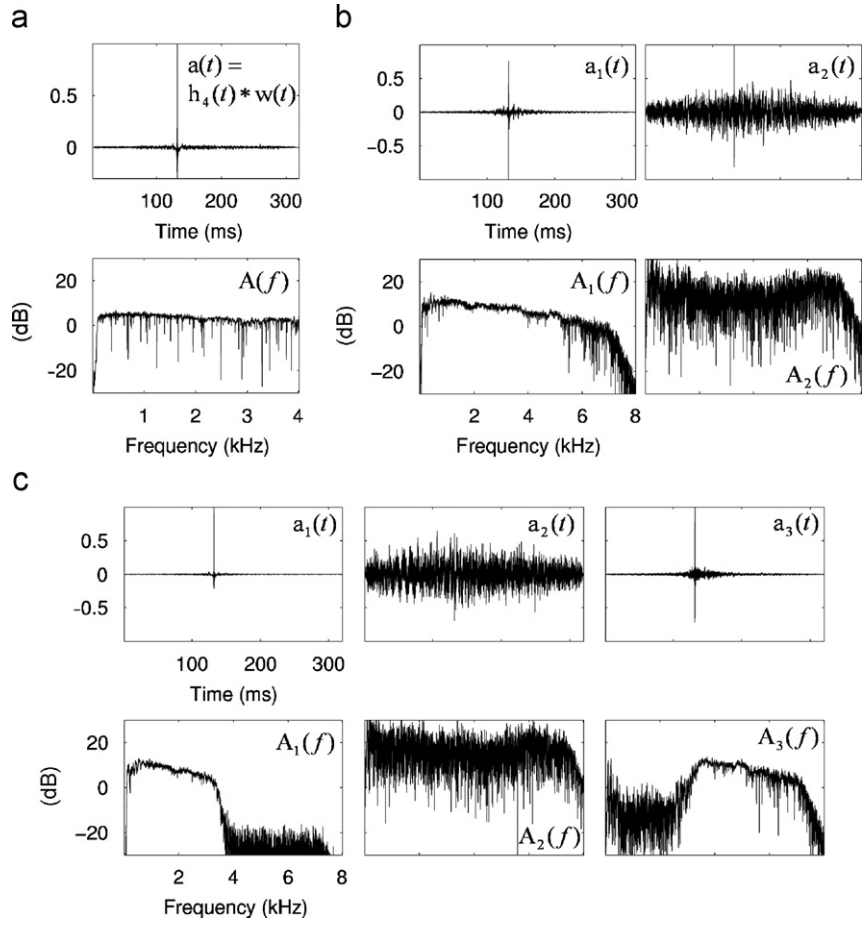



**Fig. 3.** Total convolution channel responses after applying the: (a) monaural model, (b) binaural ($M = 2$) model with two outputs and (c) binaural ($M = 3$) model with three outputs (experimental condition: $h_3$ & $h_4$ with 15 dB SNR of AWGN).

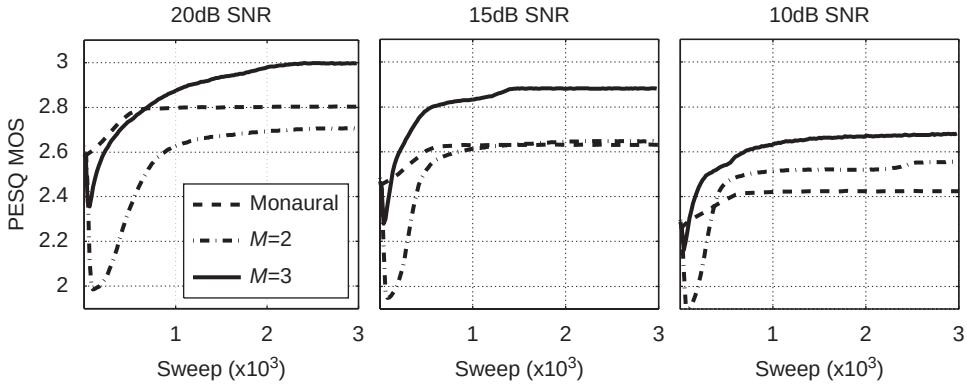



**Fig. 4.** Learning curves of PESQ MOS for the simulated reverberations (in a figure legend, '$M = 2$' and '$M = 3$' correspond to binaural models with two and three outputs, respectively).

have more taps (8192- and 16384-tap) and compared to the results of 4096-tap dereverberation filters. Updating condition of $w_{ji}$ and adaptation of $\eta$ are the same with the experiments using the simulated reverberations in Section 3. The iterative learning continued until 2000 sweeps for all three cases and the kurtosis values employed as a measure of non-Gaussianity of the output components were stabilized after the learning (i.e. assumed as a convergence).

Fig. 5 shows examples of waveforms and spectrograms of a clean speech, two recorded ones, and three output components. The resulting $u_1$, $u_2$, and $u_3$ correspond to a dereverberated high-frequency speech signal, separated measurement noise, and dereverberated low-frequency speech signal, respectively. From both waveforms and spectrograms, we can observe that the room reverberations were successfully reduced (i.e. compare s, $x_1$, $x_2$, and $u_3$).

Regarding the performance measure, it is worth to note that PESQ MOS is mainly designed for use with digital (not acoustic) interfaces to the systems under test. In this context, the PESQ MOS measure may not be adequate in our experimental setup in which only the effects of acoustic reverberations along with inherent additive noises are involved. Therefore, as more reasonable

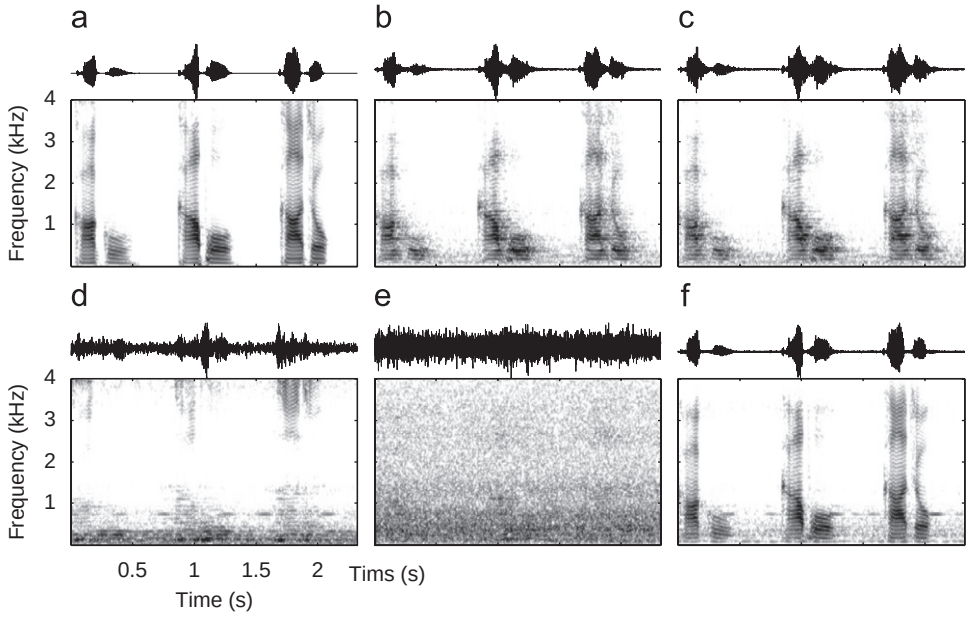



**Fig. 5.** An example of waveforms and spectrograms for real-recorded data of speaker #12 before and after applying the proposed binaural sBD model (4096-tap $w_{ji}$) with $M = 3$ (compare (a)–(c) and (f) to see the improvement of a speech quality).

**Table 1**
The word error rates (WERs) and PESQ MOS values for 13 speakers for testing as well as the averaged values with standard deviations across the speakers

| Speaker | Word error rate (%) | | | | | PESQ MOS | | | |
|---|---|---|---|---|---|---|---|---|---|
| Index | Clean | Record | 4096 | 8192 | 16 384 | Record | 4096 | 8192 | 16 384 |
| #01 | 0.0 | 10.7 | 4.0 | 1.3 | 2.7 | 2.65 | 2.96 | 3.07 | 3.06 |
| #02 | 0.0 | 13.3 | 5.3 | 1.3 | 1.3 | 2.65 | 2.86 | 3.18 | 3.15 |
| #03 | 1.3 | 23.3 | 6.7 | 4.0 | 8.0 | 2.68 | 2.96 | 3.24 | 3.15 |
| #04 | 1.3 | 19.3 | 8.0 | 4.0 | 2.7 | 2.75 | 3.12 | 3.27 | 3.13 |
| #05 | 1.3 | 14.7 | 6.7 | 4.0 | 4.0 | 2.67 | 3.02 | 3.17 | 3.13 |
| #06 | 1.3 | 12.0 | 5.3 | 2.7 | 1.3 | 2.61 | 2.66 | 3.19 | 3.07 |
| #07 | 0.0 | 12.0 | 9.3 | 4.0 | 5.3 | 2.54 | 2.67 | 3.06 | 2.97 |
| #08 | 1.3 | 11.3 | 12.0 | 2.7 | 5.3 | 2.71 | 2.76 | 3.16 | 3.13 |
| #09 | 0.0 | 7.3 | 2.7 | 0.0 | 0.0 | 2.74 | 3.02 | 3.31 | 3.27 |
| #10 | 0.0 | 6.7 | 21.3 | 10.7 | 10.7 | 2.62 | 2.51 | 2.88 | 2.86 |
| #11 | 2.7 | 24.0 | 24.0 | 12.0 | 10.7 | 2.61 | 2.58 | 3.03 | 2.87 |
| #12 | 2.7 | 18.0 | 13.3 | 5.3 | 6.7 | 2.71 | 3.00 | 3.12 | 3.04 |
| #13 | 0.0 | 8.7 | 6.7 | 1.3 | 1.3 | 2.78 | 2.96 | 3.21 | 3.19 |
| Total | $0.9 \pm 1.0$ | $13.9 \pm 5.7$ | $9.6 \pm 6.5$ | $4.1 \pm 3.5$ | $4.6 \pm 3.6$ | $2.7 \pm 0.1$ | $2.9 \pm 0.2$ | $3.2 \pm 0.1$ | $3.1 \pm 0.1$ |

Clean: original clean speech; Record: recorded speech within two microphones (averaged results); 4096, 8192, and 16 384: dereverberated speech using the corresponding number of taps for the dereverberation filters.

assessment measure, we evaluated the performance from an ASR experiment and PESQ MOS might be considered as a secondary means of a quality assessment.

A continuous density Hidden Markov Model (HMM) available in Hidden Markov Toolkit (HTK) 3.1[5] was used as a classifier. Using HTK, 39th order mel-frequency cepstral coefficients (MFCCs) including delta and acceleration coefficients (i.e. feature vector or FV) were extracted from every 25 ms of speech segments and the time difference between adjacent segments was 10 ms (i.e. temporal resolution of FVs). The resulting FVs of MFCCs from each word were used as an input of the HMM classifier (18-state left-right model with no skip). The FVs corresponding to clean isolated words (2625) from 35 speakers (the same data used for the training of whitening filter) were employed to train the model parameters of the HMM. The FVs corresponding to 75 isolated words from each of the remaining 13 speakers were then classified using the trained HMM. Error rates related to the classification of words (i.e. word error rates or WERs) were separately obtained for the clean, recorded (without dereverberation), and dereverberated speech data.

The results of WERs for each of 13 speakers' data are summarized in Table 1 along with the PESQ MOS values. Note that PESQ MOS for clean speech is 4.5. Overall, after using 8192- or 16384-tap of dereverberation filters, performances were drastically improved for virtually all speakers except only one speaker (#10) who showed degraded WER (10.7% for 8192-tap) compared to that of the recorded data (6.7%). From the retrospective analysis on this speaker's results, we found that the dereverberation filters were over-trained whereby the WER and PESQ MOS after 500 sweeps were 5.3% and 3.01, respectively. The averaged performance of both the WER and PESQ MOS across all subjects clearly

[5] http://htk.eng.cam.ac.uk/

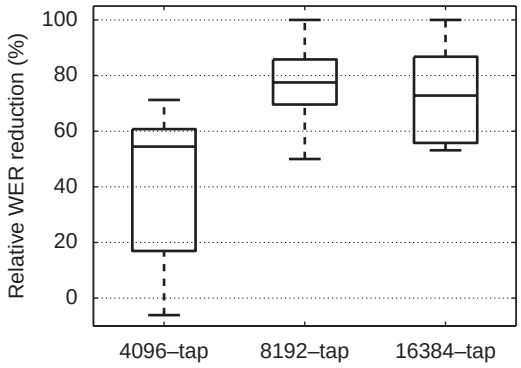



**Fig. 6.** A box and whisker plot of the relative word error rate (WER) reduction depending on the number of taps of the dereverberation filters (a box: the lower quartile, median and upper quartile values; whisker: the 5th and 95th percentile values).

reflects the efficacy of the proposed method and both measures show maximum enhancement across all speakers in the case of 8192-tap (i.e. WER: from 13.9% to 4.1%; PESQ MOS: from 2.7 to 3.2).

Fig. 6 shows the relative reduction of WER across all speakers, which is defined as

$$\text{Relative WER reduction } (\%) = \frac{\text{WER}_{\text{record}} - \text{WER}_{\text{dereverb.}}}{\text{WER}_{\text{record}}} \times 100. \tag{11}$$

Note that the relative WER reduction is saturated after using the 8192-tap filters (77.5% of median) suggesting that the inverse of the room reverberations was successfully achieved using this length of filters rather than 4096-tap of filters (54.4% of median). The slight degradation corresponding to the 16 384-tap filters (72.8% of median) indicates that the given reverberated speech data (~60 s) may not be long enough to estimate the 16 384 taps (~1 s) and this might cause under-training of the filters. Overall, these experimental results from the real recorded data suggest that the proposed model is applicable to the real room recording environment.

## 5. Conclusions

In order to improve a dereverberation performance under noisy environment, this letter proposed the binaural sBD model adopted from a BSS model. Its learning algorithm was derived from the BLS cost function in the frequency domain and experimental results were obtained from the real recording experiment as well as the simulated conditions. For various simulated conditions, the proposed model resulted better speech quality than the conventional monaural model. Also, the results of the real recorded data may indicate the feasibility of the proposed model as a pre-processing stage for real applications.

## Acknowledgments

The authors would like to thank the anonymous reviewers for their criticisms which improve this letter very much. Also, the authors thank Drs. Doh-Suk Kim and Hyung-Min Park for their fruitful discussions. This work was supported by the Brain Neuroinformatics Research Program sponsored by Korean Ministry of Commerce, Industry, and Energy.

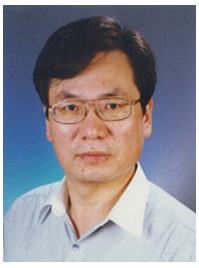
## References

[1] S. Bellini, Bussgang techniques for blind deconvolution and equalization, in: S. Haykin (Ed.), Blind Deconvolution, Prentice–Hall, Englewood Cliffs, NJ, 1994, pp. 8–52.
[2] L. De Lathauwer, A. de Baynast, Blind deconvolution of DS-CDMA signals by means of decomposition in rank-(1,L,L) terms, IEEE Trans. Signal Process. 56 (2008) 1562–1571.
[3] L. DiPersia, M. Yanagida, H.L. Rufiner, D. Milone, Objective quality evaluation in blind source separation for speech recognition in a real room, Signal Process. 87 (2007) 1951–1965.
[4] S.C. Douglas, H. Sawada, S. Makino, Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters, IEEE Trans. Speech Audio Process. 13 (2005) 92–104.
[5] J.R. Hopgood, P.J.W. Rayner, Blind single channel deconvolution using nonstationary signal processing, IEEE Trans. Speech Audio Process. 11 (2003) 476–488.
[6] Y. Huanga, J. Benestyb, J. Chena, Identification of acoustic MIMO systems: challenges and opportunities, Signal Process. 86 (2006) 1278–1295.
[7] A. Hyvarinen, Some extensions of score matching, Comput. Stat. Data Anal. 51 (2007) 2499–2512.
[8] K. Kokkinakis, A.K. Nandi, Multichannel blind deconvolution for source separation in convolutive mixtures of speech, IEEE Trans. Audio Speech Language Process. 14 (2006) 200–212.
[9] R.H. Lambert, A.J. Bell, Blind separation of multiple speakers in a multipath environment, in: Proceedings of the IEEE International Conference Acoustics Speech Signal Processing, vol. 1, 1997, pp. 423–426.
[10] M.S. Lewicki, Efficient coding of natural sounds, Nat. Neurosci. 5 (4) (2002) 356–363.
[11] C.J. Lin, Projected gradient methods for nonnegative matrix factorization, Neural Comput. 19 (2007) 2756–2779.
[12] O.L. Mangasarian, J.B. Rosen, M.E. Thompson, Convex kernel underestimation of functions with multiple local minima, Comput. Optim. Appl. 34 (2006) 35–45.
[13] T. Nakatani, K. Kinoshita, M. Miyoshi, Harmonicity-based blind dereverberation for single-channel speech signals, IEEE Trans. Audio Speech Language Process. 15 (2007) 80–95.
[14] G. Panci, S. Colonnese, P. Campisi, G. Scarano, Blind equalization for correlated input symbols: a Bussgang approach, IEEE Trans. Signal Process. 53 (2005) 1860–1869.
[15] L. Rabiner, B.H. Juang (Eds.), Fundamentals of Speech Recognition, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[16] J. Werfel, X. Xie, H.S. Seung, Learning curves for stochastic gradient descent in linear feedforward networks, Neural Comput. 17 (2005) 2699–2718.
[17] L. Zhang, A. Cichocki, S. Amari, Multichannel blind deconvolution of nonminimum-phase systems using filter decomposition, IEEE Trans. Signal Process. 52 (2004) 1430–1442.

**Jong-Hwan Lee** received his B.S. degree in Electronics Engineering from Yonsei University, Seoul, Korea, in 1998, the M.S. and Ph.D. degrees in Electrical Engineering and Computer Science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2000 and 2005, respectively. He was a visiting scholar at the Institute for Neural Computation, University California at San Diego, in 2000. He was a postdoctoral researcher at the Brain Science Research Center (BSRC) and Department of BioSystems at KAIST, in 2005. From 2005 to 2008, he was a research fellow at the Department of Radiology, Brigham and Women's Hospital and a research associate at the Harvard Medical School, Boston, MA. From 2008, he is an Instructor in Radiology, Brigham and Women's Hospital, Harvard Medical School. His research interests include machine learning algorithms and their applications in speech, image, and biomedical data. Recently, he has also worked on biomedical signal analyses such as an automated real-time registration of function MRI (fMRI) data, independent vector analysis (IVA) for group fMRI data processing, and simultaneous EEG-fMRI data analysis.

**Sang-Hoon Oh** received his B.S. and M.S. degrees in Electronics Engineering from Pusan National University in 1986 and 1988, respectively. He received his Ph.D. degree in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1999. From 1988 to 1989, he worked for the LG Semiconductor, Ltd., Korea. From 1990 to 1998, he was a senior researcher in Electronics and Telecommunications Research Institute (ETRI), Korea. From 1999 to 2000, he was with Brain Science Research Center, KAIST. In 2000, he was with Brain Science Institute, RIKEN in Japan. In 2001, he was an R&D manager of Extell Technology Corporation, Korea. Since 2002, he has been with the Department of Information Communication Engineering, Mokwon University, Daejeon, Korea. Also, he is a visiting scholar in the College of Computing, Georgia Institute of Technology, from 2008 to 2009. His research

interests are supervised/unsupervised learning for intelligent information processing, speech processing, and pattern recognition.



**Soo-Young Lee** received the B.S. degree from Seoul National University, Seoul, Korea, in 1975, the M.S. degree from the Korea Advanced Institute of Science, Daejon, Korea, in 1977, and the Ph.D. degree from the Polytechnic Institute of New York in 1984. From 1977 to 1980, he was with the Taihan Engineering Company, Seoul. From 1982 to 1985, he was also with the General Physics Corporation, Columbia, MD. In early 1986, he joined the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), as an Assistant Professor and is now a Full Professor in the Department of BioSystems and also the Department of Electrical Engineering and Computer Science. In 1997, he established the Brain Science Research Center, which is the main research organization for the Korean Brain Neuroinformatics Research Program. The research program is one of the Korean Brain Research Promotion Initiatives sponsored by the Korean Ministry of Science and Technology from 1998 to 2008, and currently about 35 Ph.D. researchers have joined the research program from many Korean universities.

Dr. Lee is a Past-President of Asia-Pacific Neural Network Assembly, and has contributed to the International Conference on Neural Information Processing as Conference Chair (2000), Conference Vice Co-Chair (2003), and Program Co-Chair (1994, 2002). He is the Editor-in-Chief of the newly established online/offline journal with a double-blind review process, Neural Information Processing–Letters and Reviews, and is on the Editorial Board for two international journals, Neural Processing Letters and Neurocomputing. He received the Leadership Award and Presidential Award from the International Neural Network Society in 1994 and 2001, respectively, and the APPNA Service Award in 2004.

His research interests have resided in artificial brain, the human-like intelligent systems based on biological information processing mechanism in our brain. He has worked on the auditory models from the cochlea to the auditory cortex for noisy speech processing, information-theoretic binaural processing models for sound localization and speech enhancement, the unsupervised pro-active developmental models of human knowledge with multi-modal man-machine interactions, and the top-down selective attention models for superimposed pattern recognitions. His research scope covers the mathematical models, neuromorphic chips, and real-world applications. Especially, he had developed a System-on-Chip (SoC) for speech recognition based on his auditory model, and a digital chip for active noise canceling and blind signal separation based on independent component analysis. Also, he has recently extended his research into brain–computer interfaces using simultaneous fMRI and EEG measurements.