Acoustic Interference Cancellation for a Voice-driven Interface in Smart TVs

Jeong-Sik Park, Gil-Jin Jang, *Member*, IEEE, Ji-Hwan Kim, *Member*, IEEE, Sang-Hoon Kim

Abstract — *A novel method is proposed to improve the voice recognition performance by suppressing acoustic interferences that add nonlinear distortion to a target recording signal when received by the recognition device. The proposed method is expected to provide the best performance in smart TV environments, where a remote control collects command speech by the internal microphone and performs automatic voice recognition, and the secondary microphone equipped in a TV set provides the reference signal for the background noise source. Due to the transmission channel, the original interference is corrupted nonlinearly, and the conventional speech enhancement techniques such as beamforming and blind signal separation are not applicable. The proposed method first equalizes the interference in the two microphones by maximizing the instantaneous correlation between the nonlinearly related target recording and reference signal, and suppresses the equalized interference. To obtain an optimal estimation of the equalization filter, a method for detecting instantaneous activity of interference is also proposed. The validity of the proposed method is proved by the improvement in automatic voice recognition performance in a simulated TV room where loud TV sounds or babbling speech interfere in a user's commanding speech*[1].

Index Terms — Acoustic interference cancellation, Speech enhancement, Voice recognition, Smart TV interface.

I. Introduction

Automatic voice recognition has been steadily advanced with an absolute necessity for human-machine interaction [1], [2]. Recently, a number of consumer electronic devices including smartphones and smart TVs have become powerful enough to be equipped with voice recognition capability [3]-[5]. Especially in smart TVs with many novel functions such as internet access or contents search, voice-driven interface is able to provide plenty of convenience by delivering commands and keywords via voice.

An example of the voice-driven TV interfaces is illustrated in Fig. 1. The user's voice is recorded by a microphone in a remote control device, and the automatic voice recognition is carried out to understand and respond to the user intention. Although voice recognition is an essential and primary task in the user interface, the recognition system is inevitably exposed to loud sounds from TV speakers and other background noises that seriously interfere with the recognition process. Such types of interferences are classified to "non-stationary noises" because their frequency characteristics change considerably over time. When non-stationary noises are added, the target voice is contaminated non-uniformly across frequency and time, and voice recognition performance is significantly degraded. Hence, cancellation of such an acoustic interference is a great challenge for the successful implementation of the voice-driven interface in smart TVs.

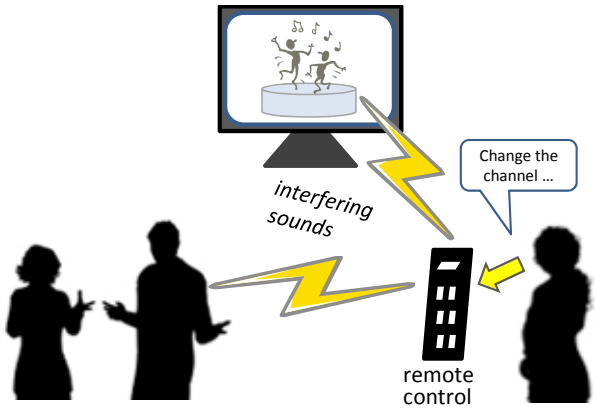

Fig. 1. Illustration of voice-driven interface in a typical smart TV environment.

This paper is organized as follows. Section II introduces the conventional methods for non-stationary noise cancellation. Section III describes the proposed method in detail. Section IV demonstrates the simulated experimental setups and the results. Finally, conclusions are presented in Section V.

II. Conventional Approaches for Non-stationary Noise Cancellation

Conventional noise cancellation methods aim at eliminating stationary noises; consequently, their performance with

[1] This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (no. 2012-0008090 and no. 2011-0005419), and by the Ministry of Knowledge Economy, Korea.

Jeong-Sik Park is with the Department of Intelligent Robot Engineering, Mokwon University, Daejeon, 302-729, South Korea (e-mail: parkjs@mokwon.ac.kr).

Gil-Jin Jang is with the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 689-798, South Korea (e-mail: gjang@unist.ac.kr).

Ji-Hwan Kim is with the Department of Computer Science and Engineering, Sogang University, Seoul, 121-742, South Korea (e-mail: kimjihwan@sogang.ac.kr).

Sang-Hoon Kim is with the Research Department of Spoken Language Processing Section of the Electronics and Telecommunications Research Institute (ETRI), Daejeon, 305-700, South Korea (e-mail: ksh@etri.re.kr).

Contributed Paper
Manuscript received 12/31/12
Current version published 03/25/13
Electronic version published 03/25/13.

speech-like noises is very poor [5], [6]. To tackle the problem of non-stationary noise cancellation, many researchers have proposed various techniques using multiple microphones [7]-[13]. Among them, a family of algorithms utilizing the geometric information is called beamforming [7]-[9], and another family of algorithms that does not require any information of the source characteristics and microphone locations is called Blind Signal Separation (BSS) [10]-[13]. These methods exploit inverse filters canceling out non-stationary interfering sounds coming from specific directions. However, there are a number of requirements to be satisfied:

1. The number of microphones should be equal to or more than the number of sound sources.
2. The position of microphone relative to the source is fixed or slowly changing.
3. The microphone signals should be recorded simultaneously to find the exact delay.

These requirements severely degrade performance in practical situations. There are usually many interfering sounds, and beamforming or BSS algorithms fail to suppress the interference arising from the first requirement. The third requirement is not always satisfied because of the cost involved in the synchronization of multiple recordings.

This paper proposes an algorithm for canceling the interfering sounds with stereo recordings of the nonlinear mixing of more than two sources. The proposed method is based on Adaptive Noise Cancellation (ANC) [14]. An assumption is made that the primary recording is a mixture of the target sound and several interfering signals, and the secondary recording called the reference is a mixture of only interfering signals.

Our proposed algorithm is particularly applicable for smart TV environments in which two individual devices, a remote control and a TV set, are equipped with microphones. The primary microphone that is in the remote control directly receives noise-contaminated speech from a target speaker to perform automatic voice recognition while the secondary microphone in the TV collects the background noise and provides the reference signal for the background noise source. Further details of this approach are described in Section IV.

III. The Proposed Noise Cancellation Method

This chapter explains the proposed interference equalization and target activity detection algorithms in detail.

A. Interference Equalization

The proposed method assumes that the primary microphone equipped in a remote control picks up a mixture of the target sound and interference, and the signal recorded by the secondary microphone in a TV set contains interference only. This is expressed as

$$x_0(t) = s(t) + r_0(t), \quad x_1(t) = r_1(t), \tag{1}$$

where $x_0(t)$ and $x_1(t)$ are the signals recorded by the primary and secondary microphones, $s(t)$ is the target source, $r_0(t)$ is the interference mixed in the primary recording, and $r_1(t)$ is the interference picked up by the secondary microphone. Because the assumed mixing architecture of ANC is linear, it may not model the linear filter's mismatch to the nonlinear, realistic conditions [15], [16]. Using short-time Fourier transform, (1) is rewritten as

$$X_0(\omega) = S(\omega) + R_0(\omega), \quad X_1(\omega) = R_1(\omega), \tag{2}$$

where ω is a frequency variable, complex values $X_0(\omega)$, $X_1(\omega)$, $S(\omega)$, $R_0(\omega)$, and $R_1(\omega)$ represent the DFTs of $x_0(t)$, $x_1(t)$, $s(t)$, $r_0(t)$, and $r_1(t)$, respectively. We assume that $R_0(\omega)$ can be approximated by a scalar multiple of $R_1(\omega)$, so that the unknown target source $S(\omega)$ should be recovered by

$$\begin{aligned} \hat{S}(\omega) &= X_0(\omega) - B(\omega)X_1(\omega) \\ &\approx S(\omega) + \left(R_0(\omega) - B(\omega)R_1(\omega)\right), \end{aligned} \tag{3}$$

where $B(\omega)$ is a constant that equalizes $R_0(\omega)$ and $R_1(\omega)$. The estimate of the target source $\hat{S}(\omega)$ becomes the true one if and only if $R_0(\omega) = B(\omega)R_1(\omega)$, i.e., when a perfect equalization is made. Fig. 2 shows an overview of the proposed method. The equalized reference signal is subtracted from the input to generate the estimate of the target. To obtain a phase-independent error criterion, we enforce $B(\omega)$ to be a positive real number, and use a squared difference of the Power Spectral Densities (PSDs) between the interferences in the primary recording and the equalized secondary recording:

$$E(\omega) = \left|R_0(\omega) - B(\omega)R_1(\omega)\right|^2. \tag{4}$$

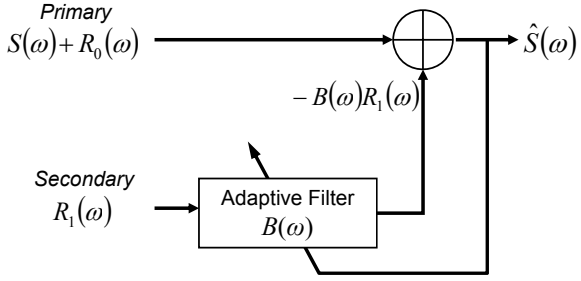

Fig. 2. Block diagram of the proposed method.

An objective function $J(\omega)$ is defined by the sum of the equalization errors in (4) over all of the short time analysis frames, expressed as

$$\begin{aligned} J(\omega) &= \sum_{all\ n} E(\omega, n) \\ &= \sum_{all\ n} \left|R_0(\omega, n) - B(\omega)R_1(\omega, n)\right|^2 \end{aligned} \tag{5}$$

where n is the frame index. To find an optimal equalization filter, we differentiate $J(\omega)$ with respect to $B(\omega)$:

$$\begin{aligned}\frac{\partial J(\omega)}{\partial B(\omega)} &= \sum_{all\ n} -2\left(\overline{R_0(\omega,n)} - \overline{B(\omega)R_1(\omega,n)}\right)R_1(\omega,n) \\ &= -2\sum_{all\ n}\left[\overline{R_0(\omega,n)}R_1(\omega,n) - \overline{B(\omega)}\,\overline{R_1(\omega,n)}R_1(\omega,n)\right] \\ &= -2\left[\sum_{all\ n}\overline{R_0(\omega,n)}R_1(\omega,n) - \overline{B(\omega)}\sum_{all\ n}\left|R_1(\omega,n)\right|^2\right],\end{aligned} \tag{6}$$

where $\overline{(\cdot)}$ is a complex conjugate operator. A closed-form solution to the minimization of $J(\omega)$ is obtained by equating the derivative to zero:

$$\begin{aligned}B^*(\omega) &= \frac{\sum_{all\ n} R_0(\omega,n)\overline{R_1(\omega,n)}}{\sum_{all\ n}\left|R_1(\omega,n)\right|^2} \\ &\approx \frac{\sum_{all\ n}\left|R_0(\omega,n)R_1(\omega,n)\right|}{\sum_{all\ n}\left|R_1(\omega,n)\right|^2}.\end{aligned} \tag{7}$$

The approximation in the second line of (7) is to enforce the value of $B(\omega)$ to be positive-real. This in turn implies that only the PSDs of the two channels are considered, and their phase difference is ignored.

B. Target Activity Detection and Iterative Adaptation of the Equalization Filter

In (2), because the primary input is a mixture of the target and the interference, $R_0(\omega,n)$ is not available. Therefore we employ a method for detecting the activity of the target source in the primary recording, so that the target is excluded from computing the equalization filter. In all of the short-time analysis frames at frequency ω, they are classified into the following two classes [16]:

$$\begin{aligned}C_{S+R}(\omega) &= \left\{n \mid X_0(\omega,n) = S(\omega,n) + R_0(\omega,n)\right\} \\ C_R(\omega) &= \left\{n \mid X_0(\omega,n) = R_0(\omega,n)\right\}\end{aligned} \tag{8}$$

where n is the frame index, $C_R(\omega)$ is a set of indexes at which the primary recording contains interference only, and $C_{S+R}(\omega)$ is the complementary of $C_R(\omega)$. For a given equalization factor $B(\omega)$, we assume that the PSD of the interference in the primary recording is equal to or smaller than the PSD of the estimated interference from the secondary recording, which is reflected in the following decision rule:

$$\begin{aligned}n \in C_{S+R}(\omega) \quad &if \quad \left|X_0(\omega,n)\right| > B(\omega)\left|X_1(\omega,n)\right| \\ n \in C_R(\omega) \quad &if \quad \left|X_0(\omega,n)\right| \le B(\omega)\left|X_1(\omega,n)\right|.\end{aligned} \tag{9}$$

The equalization filter is updated using only the frames whose indexes are in $C_R(\omega)$, such that

$$\begin{aligned}B\left(\omega \mid C_R(\omega)\right) &= \frac{\sum_{n\in C_R(\omega)}\left|R_0(\omega,n)R_1(\omega,n)\right|}{\sum_{n\in C_R(\omega)}\left|R_1(\omega,n)\right|^2} \\ &\approx \frac{\sum_{n\in C_R(\omega)}\left|X_0(\omega,n)X_1(\omega,n)\right|}{\sum_{n\in C_R(\omega)}\left|X_1(\omega,n)\right|^2}.\end{aligned} \tag{10}$$

Because the update of $B\left(\omega|C_R(\omega)\right)$ leads to the change of the index set $C_R(\omega)$, we use an Expectation-Maximization (EM) manner [17] in deriving the following algorithm:

- Input: observed signals $X_0(\omega,n)$, $X_1(\omega,n)$
- Output: equalization filter $B^*(\omega)$

1. Initialize $C_R = \{1,2,\ldots,N\}$ where N is the total number of frames
2. Update $B\left(\omega|C_R\right)$ with respect to the current C_R by
   $B\left(\omega|C_R\right) = \sum_{n\in C_R}\left|X_0(\omega,n)X_1(\omega,n)\right| \Big/ \sum_{n\in C_R}\left|X_1(\omega,n)\right|^2$
3. Update the index set C_R by
   $C_R^{new} = \left\{n \mid \left|X_0(\omega,n)\right| \le B\left(\omega|C_R\right)\left|X_1(\omega,n)\right|\right\}$
4. Repeat steps 2 and 3 until there is no change in the index set C_R
5. Assign the final value of $B\left(\omega|C_R\right)$ as the optimal Wiener filter coefficient at frequency ω: $B^*(\omega) = B\left(\omega|C_R^{final}\right)$

We omit the frequency index ω from C_R for a compact notation. In Step 1, all the frames are assigned to C_R because $B(\omega)$ is not known initially.

C. Wiener Filter Derivation

Once the equalization filter coefficients $B^*(\omega)$ are obtained, the amount of interference in the primary channel recording at frequency ω is approximated by $B^*(\omega)X_1(\omega)$. A Wiener filter at frame n and frequency ω suppressing the interference estimate from the PSD of the primary channel is derived by

$$\begin{aligned}W(\omega,n) &= \max\left[\varepsilon, \frac{\left|X_0(\omega,n)\right|^2 - \left|B^*(\omega)X_1(\omega,n)\right|^2}{\left|X_0(\omega,n)\right|^2}\right] \\ &= \max\left[\varepsilon, 1 - \frac{\left|B^*(\omega)X_1(\omega,n)\right|^2}{\left|X_0(\omega,n)\right|^2}\right],\end{aligned} \tag{11}$$

where a hard-limiting function $\max[\varepsilon,\cdot]$ ensures that the computed Wiener filter gain should always be higher than a certain limit defined by a nonnegative constant ε. In our experiments, $\varepsilon = 0.15$ (-16 dB) provided the best tradeoff between the prevention of musical noise and maintaining a decent voice recognition performance. $W(\omega,n)$ is then converted to minimum-phase, time-domain FIR filter and

applied to the time-domain primary input, $x_0(t)$ in (1), and overlap-added with trapezoidal windowing. The advantage of the Wiener filter is that it effectively absorbs the small amount of mismatch between the actual filter and the estimated one to prevent musical noise interference [15].

IV. Experimental Results

Comparing Signal-to-Noise Ratio (SNR) of noise-interfered sound and that of noise-suppressed sound is a useful method to verify the efficiency of a noise cancellation approach. However, our proposed method targets at the nonlinear mixing of the original sound and noise signals, and an estimation of SNR is unfeasible because the original sound is not provided. Alternatively, the validity of the proposed method is verified by automatic voice recognition experiments.

We conducted voice recognition experiments using the Speech Separation Challenge (SSC) database [18]. This database is designed for assessing the effectiveness of a noise cancellation algorithm in a simple voice recognition task. Several participants uttered short sentences comprising of exactly 6 words, using a *command-color-preposition-letter-number-adverb* format. For example, "bin blue at F 2 now." We considered this command format reflects typical voice command forms in smart TV controls. The database has a training set, which consists of 17,000 utterances uttered by 34 speakers. The Hidden Markov Models (HMMs) were obtained using the Hidden Markov model Toolkit (HTK) [19], as suggested by SSC coordinators. The adopted features were 12 Mel-Frequency Cepstral Coefficients (MFCCs) plus log energy, plus their velocities and accelerations, resulting in a 39-dimensional vector at every 10 ms [18]-[20]. A separate testing set of 600 utterances was also provided. There are no overlaps between the training and the test data. All data were recorded in a quiet environment; i.e., without any background noise.

The analysis settings of the proposed noise cancellation method are: sampling frequency 8 kHz, shift size 10 ms (80 samples), analysis frame length 20 ms (160 samples), and hamming windowing. In re-synthesis, time-domain filters of order 48 are derived from the Wiener filters in (11), and applied to the input frames. The resulting frames are overlap-added by trapezoidal windows of 24-sample overlaps between the neighboring frames.

A. Setup of Recording Environment

To simulate the voice recognition in a TV room, a recording environment was set up as shown in Fig. 3. The primary recording was collected by a condenser microphone connected to a computer simulating the remote control, considering that users' voice commands are submitted to a voice recognition module in a remote control of a smart TV. In addition, the secondary recording was obtained by an internal microphone of another laptop computer, simulating the internal microphone in the smart TV. The two microphones were located 1 m apart as a typical distance between a viewer and a TV. Test data files were played at a distance of 0.3 m from the primary microphone; thus, the distance between the sound source and the secondary microphone was 1.3 m. Meanwhile, several types of interfering sounds provided by the AURORA database [20] were played far from both of the two microphones in order to simulate a general situation for noise sounds in a TV room. We varied the SNR of the voice signal to the interfering signal from 20 dB to 6 dB to investigate the performance variation of the proposed method based on the interference level. Because two microphone signals were recorded by independent computers, they were not perfectly synchronized in time, and nonlinear distortion from the PC's sound card was added to the recorded signal. The start time was matched by maximizing the correlation between the two recordings.

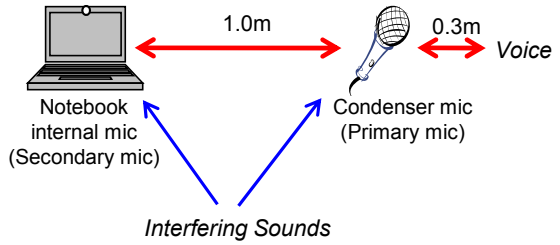

Fig. 3. Recording setup for the simulated TV room environment.

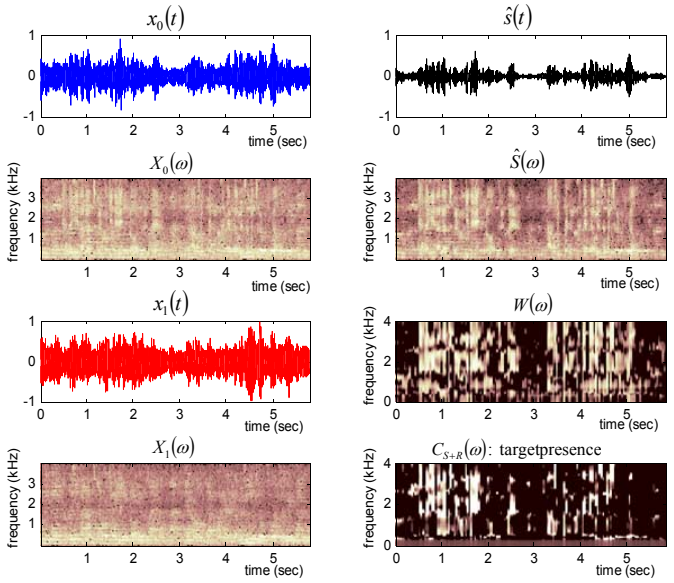

Fig. 4. Interference cancellation result.

B. Evaluation Results

First, we analyzed several waveform and spectrogram figures for a noise-contaminated speech data and noise-suppressed data. Fig. 4 demonstrates the result. For an input SNR 6 dB, the left four graphs are $x_0(t)$, $X_0(\omega)=S(\omega)+R_0(\omega)$, $x_1(t)$, and $X_1(\omega)=R_1(\omega)$, respectively. The top two graphs $\hat{s}(t)$ and $\hat{S}(\omega)$ on the right side are the estimate of the target source obtained by the proposed interference cancellation method. The bottom two graphs are the estimated Wiener filters and the visualization of $C_{S+R}(\omega)$, roughly showing target presence

probability. As shown in these figures, our proposed method successfully suppressed the interfering sounds from noise-contaminated speech while the Wiener filter coefficients tracked the original target speech quite well.

TABLE I
VOICE RECOGNITION PERFORMANCE COMPARISON OF THE PROPOSED METHOD UNDER VARIOUS SNR CONDITIONS

Recording SNR	20 dB	12 dB	6 dB
None	95.6%	84.5%	64.6%
Proposed	95.3%	89.1%	72.8%

Next, the results of voice recognition experiments are summarized in TABLE I. The proposed method aims at canceling the interfering sounds with stereo recordings of the nonlinear mixing of more than two sources. Because the conventional ANC or BSS techniques assume that the mixing filter is linear, separation of the nonlinearly mixed signals is not achieved with those methods. For this reason, we just investigated the voice recognition accuracy of the noise-interfered speech (None) and that of the noise-suppressed speech (Proposed). For 20 dB SNR, the proposed method was on a par with, or slightly better than None (no processing). In the other SNR conditions, our proposed method exhibited more than 4% improvement at 12 dB, and 8% improvement at 6 dB. As the SNR decreased, greater improvement was obtained. Deployment of a speech recognition system is not practical for SNRs lower than 6 dB due to the severe speech distortion caused by interference. For a fair evaluation, all HMMs are trained using the original clean speech. The proposed method significantly improved the voice recognition performance in the presence of interference, without diminishing performance in the clean condition.

V. CONCLUSIONS

We proposed a novel method to reduce non-stationary acoustic noise added to a speech signal recorded by dual independent devices. The method masks power spectral densities of the input signal to suppress the estimated interference using another device's recording as a reference. An algorithm for equalizing the difference between the two recordings was proposed. The proposed method has an advantage over the ordinary BSS and beamforming algorithms in that it can handle non-stationary noise sounds. In particular, it can be applied to more than two sources of interference, as in smart TV environments surrounded by loud interfering sounds. To verify the efficiency of our approach in the voice-driven TV interface, we conducted voice recognition experiments in a simulated TV room, using two microphone devices as a remote control and a TV set. These experiments and their evaluation results give a new idea about the acoustic interference cancellation in smart TV environments, and also a strong possibility of applying our approach to the voice-driven interface of smart TVs.

Future work includes applying the proposed method to more realistic environments and improving the performance in those conditions.

REFERENCES

[1] R. Kil and Y. Kim, "Zero-crossing-based speech segregation and recognition for humanoid robots," *IEEE Trans. Consumer Electron.,* vol. 55, no. 4, pp. 2341-2348, Nov. 2009.

[2] J. Park, G. Jang, and Y. Seo, "Music-aided affective interaction between human and service robot," *EURASIP J. Audio Speech Music Process.*, vol. 2012, no. 5, pp. 1-13, Jan. 2012.

[3] S. Chang, D. Yook, and Y. Kim, "A voice trigger system using keyword and speaker recognition for mobile devices," *IEEE Trans. Consumer Electron.,* vol. 55, no. 4, pp. 2377-2384, Nov. 2009.

[4] J. Park, G. Jang, and J. Kim, "Multistage utterance verification for keyword recognition-based online spoken content retrieval," *IEEE Trans. Consumer Electron.*, vol. 58, no. 3, pp. 1000-1005, Aug. 2012.

[5] G. Jang, J. Park, J. Kim, and Y. Seo, "Line spectral frequency-based noise suppression for speech-centric interface of smart devices," *Adv. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3-8. Nov. 2011.

[6] G. Jang and H. Cho, "Efficient spectrum estimation of noise using line spectral pairs for robust speech recognition," *Electron. Lett.*, vol. 47, no. 25, pp. 1399-1401, Dec. 2011.

[7] M. Z. Ikram, D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," *Proc. of the IEEE Int. Conf. on Acous., Speech, and Sig. Proc.*, pp. 881-884, May 2002.

[8] J. C. Chen, K. Yao, and R. E. Hudson, "Acoustic source localization and beamforming: theory and practice," *EURASIP J. Applied Sig. Proc.*, vol. 2003, no. 4, pp. 359-370. Mar. 2003.

[9] J. Hong, S. Jeong, and M. Hahn, "Robust GSC-based speech enhancement for human machine interface," *IEEE Trans. Consumer Electron.,* vol. 56, no. 2, pp. 965-970, May 2010.

[10] I. Lee and G. Jang, "Independent vector analysis using densities represented by chain-like overlapped cliques in graphical models for separation of convolutedly mixed signals," *Electron. Lett.,* vol. 45, no. 13, pp. 710-711, Jun. 2009.

[11] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech and Aud. Proc.,* vol. 8, no. 3, pp. 320-327, May 2000.

[12] T. Lee, A. Bell, and R. Lambert "Blind separation of delayed and convolved sources," *Adv. in Neur. Inf. Proc. Sys.,* vol. 9, pp. 758-764, 1997.

[13] I. Lee and G. Jang, "Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation," *EURASIP J. Adv. in Sig. Proc.,* vol. 2012, no. 113, pp. 1-12. May 2012.

[14] B. Widrow, J. Glover, J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Zeidler, E. Dong, and R. Goodlin, "Adaptive noise cancelling: principles and applications," Proc. of the IEEE, vol. 63, no. 12, pp. 1692-1716, Dec. 1975.

[15] S. Boll. "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acous., Speech and Sig. Proc.,* vol. 27, no. 2, pp. 113-120, Apr. 1979.

[16] N. Kim and J. Chang, "Spectral enhancement based on global soft decision," *IEEE Sig. Proc. Lett.,* vol. 7, no. 5, pp. 108–110, May 2000.

[17] M. Sato and S. Ishii, "On-line EM algorithm for the normalized Gaussian network," *Neural comp.,* vol. 12, no. 2, pp. 407-432, Feb. 2000.

[18] M. Cooke and T. Lee, "Speech separation challenge," *Proc. of INTERSPEECH,* Sep. 2006.

[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *Hidden Markov model toolkit (HTK); version 3.4*, Cambridge University Engineering Department, 2006.

[20] D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," *Proc. of Int. Conf. on Spoken Lang. Proc.,* Oct. 2000.

BIOGRAPHIES

Jeong-Sik Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. and Ph.D. degree in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2010, respectively. From 2010 to 2011, he was a Post-Doc. researcher in the Computer Science Department, KAIST. He is now an assistant professor in the Department of Intelligent Robot Engineering, Mokwon University. His research interests include speech emotion recognition, speech recognition, speech enhancement, and voice interface for human-computer interaction.

Gil-Jin Jang (M'10) is an assistant professor at Ulsan National Institute of Science and Technology (UNIST), South Korea. He received his B.S., M.S., and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, South Korea in 1997, 1999, and 2004, respectively. From 2004 to 2006 he was a research staff at Samsung Advanced Institute of Technology and from 2006 to 2007 he worked as a research engineer at Softmax, Inc. in San Diego. From 2008 to 2009 he joined Hamilton Glaucoma center at University of California, San Diego as a postdoctoral employee. His research interests include acoustic signal processing, pattern recognition, speech recognition and enhancement, and biomedical signal engineering.

Ji-Hwan Kim (M'09) received the B.E. and M.E. degrees in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 1996 and 1998 respectively and Ph.D. degree in Engineering from the University of Cambridge in 2001. From 2001 to 2007, he was a chief research engineer and a senior research engineer in LG Electronics Institute of Technology, where he was engaged in development of speech recognizers for mobile devices. In 2005, he was a visiting scientist in MIT Media Lab. Since 2007, he has been a faculty member in the Department of Computer Science and Engineering, Sogang University. Currently, he is an associate professor. His research interests include spoken multimedia content search, speech recognition for embedded systems and dialogue understanding.

Sang-Hun Kim received the BS in electrical engineering from Yonsei University, Seoul, Korea, in 1990, and the MS degree in electrical and electronic engineering from KAIST, Daejeon, Korea, in 1992. He received his PhD from the Department of Electrical, Electronic, Information, and Communication Engineering at the University of Tokyo, Japan, in 2003. Since 1992, he has been with the Research Department of Spoken Language Processing Section of ETRI, Daejeon, Korea. Currently, he is a principal researcher in the Automatic Speech Translation Research Team. His interests include speech synthesis, speech recognition, and speech signal processing.