Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 10(2): 162-167, Jun. 2012 **Regular Paper**

Speech Processing System Using a Noise Reduction Neural Network Based on FFT Spectrums

Jae-Seung Choi, ***Member, KIICE***

Department of Electronic Engineering, Silla University, Busan 617-736, Korea

Abstract

This paper proposes a speech processing system based on a model of the human auditory system and a noise reduction neural network with fast Fourier transform (FFT) amplitude and phase spectrums for noise reduction under background noise environments. The proposed system reduces noise signals by using the proposed neural network based on FFT amplitude spectrums and phase spectrums, then implements auditory processing frame by frame after detecting voiced and transitional sections for each frame. The results of the proposed system are compared with the results of a conventional spectral subtraction method and minimum mean-square error log-spectral amplitude estimator at different noise levels. The effectiveness of the proposed system is experimentally confirmed based on measuring the signal-to-noise ratio (SNR). In this experiment, the maximal improvement in the output SNR values with the proposed method is approximately 11.5 dB better for car noise, and 11.0 dB better for street noise, when compared with a conventional spectral subtraction method.

Index Terms: Speech processing, Neural network, Amplitude and phase spectrums, Background noise

I. INTRODUCTION

In the field of noise processing, background noise is one of the important research problems. In particular, it has been noted that system reliability, and especially the speech recognition ratio, can be significantly decreased by background noise [1, 2]. Background noise includes various non-stationary noises existing in the real environment such as street noise and babble noise. Thus, typical background noises cannot be simply eliminated with a Wiener filter or spectral subtraction, but require more skillful techniques. Therefore, to solve the above-mentioned problems, this paper proposes a speech processing system using an auditory system and noise reduction neural network that is effective in various noisy environments.

Several noise suppression and speech processing methods have been proposed, such as spectral subtraction [3-5], neural networks [6, 7], minimum mean-square error estimators [1, 8, 9], and a time-delay neural network [10]. The spectral subtraction method is used in signal processing to process speech adaptively according to the noise intensity, thereby enhancing the performance. For instance, in a study by Lim et al. [4], speech intelligibility was improved when choosing four different lengths of filter for the pitch period according to four different input signal-to-noise ratios (SNRs). In this paper, the parameters of lateral inhibition function are adjusted frame by frame. For details, the parameters are used to adjust the width and amplitude of the lateral inhibition filter at each frame, respectively [10, 11].

Recently, developments of auditory processing systems have been reported in various studies imitating functions of signal processing in auditory processing systems [10-12], and one of these auditory mechanisms, called lateral inhibition, is used as the speech processing system in this

Received 21 March 2012, **Revised** 23 April 2012, **Accepted** 26 April 2012
***Corresponding Author** E-mail: jschoi@silla.ac.kr

Open Access **http://dx.doi.org/10.6109/jicce.2012.10.2.162** print ISSN:2234-8255 online ISSN:2234-8883

paper. In the area of speech signal processing, a neural network (NN) is applied to speech recognition area, while in the area of speech enhancement and noise reduction, the major application of NNs is the extraction of speech sections from a noisy speech signal. Moreover, amplitude spectrums contain more information than phase spectrums when speech signals are generated by fast Fourier transform (FFT).

In this paper, the proposed system reduces the noise of a speech signal by using the proposed NN based on FFT amplitude spectrums and phase spectrums, and implements auditory processing frame by frame according to the detected results after detected voiced sections and transitional sections for each frame. Thereafter, the proposed system restores the FFT amplitude and phase spectrums using an inverse fast Fourier transform (IFFT).

To evaluate the proposed speech processing system, the SNR is measured as an indication of the speech intelligibility, and the results confirm the effectiveness of our solution in the case of white, car, restaurant, subway, and street noise.

II. MODEL OF SPEECH AND NOISE

A. Speech and Noise Database

The speech data used in the experiment are included in the Aurora-2 database (DB) that consists of English-connected digits, recorded in clean environments with a sampling frequency of 8 kHz [13]. The Aurora-2 DB offers two different training modes, i.e., clean training and multi-conditional training modes. The clean training mode includes 8,440 utterances, which contains the voices of 55 male and 55 female adult recordings. The same 8,440 utterances are also used in the multi-conditional training mode. For each training mode, three test sets (sets A, B, and C) are provided. These clean utterances are artificially contaminated by adding different types of noise (subway, babble, car, etc.) to the clean utterances at different SNR levels. In test set A, four types of noise (subway, car, exhibition hall, and babble) are added to the clean utterances at SNR levels of -5, 0, 5, 10, 15, and 20 dB, while in test set B, another set of four different types of noise (restaurant, street, airport, and train station) are added to the clean utterances at the same SNR levels. In test set C, two types of noise (subway and street) used in sets A and B are added. The three test sets are composed of 4,004 utterances from 52 male and 52 female speakers [10, 11].

B. Proposed Methods

In this paper, a weighted spectral average filter is adopted to reduce unexpected peaks at each frame, as shown in Eq. (1) [10, 11].

$$\bar{S}_A^i(\omega) = \frac{1}{5}\sum_{j=-2}^{2} W_j S_A^{(i-j)}(\omega) \tag{1}$$

In Eq. (1), the weighted values for the parameters of W_j are used as follows: $W_{-2} = 0.5$, $W_{-1} = 1.2$, $W_0 = 1.6$, $W_1 = 1.2$, and $W_2 = 0.5$, where, $\bar{S}_A^i(\omega)$ is the FFT spectral average of the i^{th} frame, $S_A^i(\omega)$ is the original speech spectrums. Moreover, this study uses the function of spectral lateral inhibition [11, 12], thereby emphasizing the spectral peaks of speech by an excitation area, while simultaneously suppressing the noise spectral valleys by two inhibition areas, as shown in Fig. 1. The parameters for $F_{i=l,e,r}$ show the amplitude of the impulse response as follows: $F_e = 1.0$, $F_l = -0.7$ and $F_r = -0.3$, where F_e represents the amplitude for an excitation area, F_l and F_r represent the amplitudes for two inhibition areas, α is used to adjust the whole amplitude for an excitation area and two inhibition areas, and the parameters for $B_{i=l,e,r}$ show the width of the lateral inhibition function. For noise reduction, the parameters are restricted to satisfy the following Eq. (2). Therefore, the noise signal is reduced because the average sum values for the weighted noise is zero [10, 11].

$$B_l F_l + B_e F_e + B_r F_r = 0 \tag{2}$$

In this paper, the parameters of W_j, $F_{i=l,e,r}$, and $B_{i=l,e,r}$ are experimentally determined to be the optimum values.

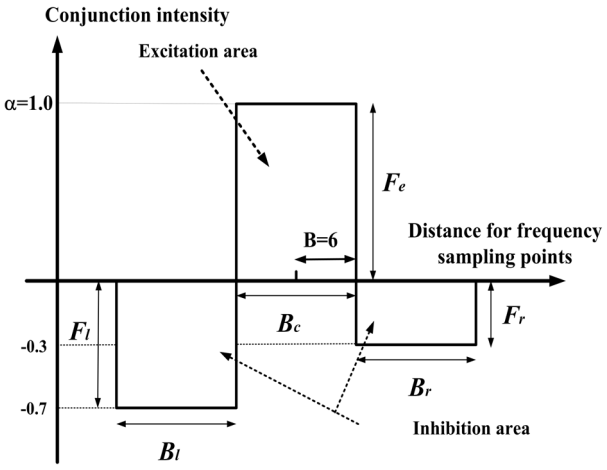

Fig. 1. Modeling of lateral inhibition function.

C. Proposed Noise Reduction NN

This paper proposes an algorithm using a noise reduction neural network (NRNN) that is constructed for the FFT amplitude and phase spectrum areas, allowing more effective correlation of the added information. In this experiment, a perceptron type NN is used and trained by a back-propagation (BP) algorithm, as shown in Fig. 2. The input and output for each unit approximates a sigmoid function, whose output range is from –1.0 to +1.0, as given by Eq. (3) [10, 11].

$$f(x_j) = \frac{2.0}{1.0 + \exp(-\sum_i w_{ji} x_i + \theta)} - 1.0 \quad (3)$$

In this equation, $f(x_j)$ is the output of unit j in the upper subnet, x_i is the output of unit i in the lower subnet, w_{ji} is the connection weight between the input unit i and the output unit j, and θ is a threshold in each unit. The proposed NN is composed of three layers and the composition of the NN is 32-64-32. In this experiment, the number of units in the hidden layer was adjusted to 64, as this produced the best convergence of the NN.

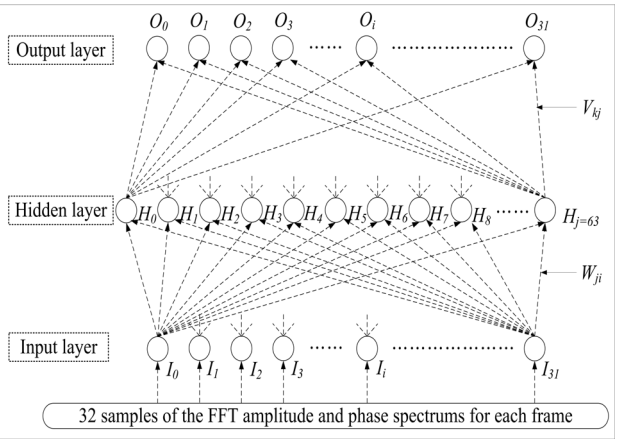

Fig. 2. The construction of a three-layer neural network. FFT: fast Fourier transform.

In order to essentially reduce the background noise, this paper proposes a noise processing algorithm using the BP training method of the NRNN system, which is composed of an NN based on amplitude and phase spectrums, as shown in Fig. 3.

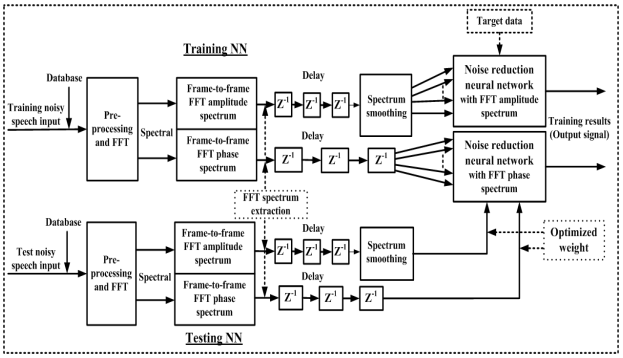

Fig. 3. Proposed noise reduction neural network system. FFT: fast Fourier transform, NN: neural network.

The proposed NN was trained using twenty types of noisy speech data selected randomly from test set A, and artificially added at five different SNRs (20 dB, 15 dB, 10 dB, 5 dB, and 0 dB) for four different noises (i.e., white, car, restaurant, and subway noise). Therefore, the proposed NN was trained using five kinds of networks: 1) input signal-to-noise ratio (SNR_{IN}) = 20 dB; 2) SNR_{IN} = 15 dB; 3) SNR_{IN} = 10 dB; 4) SNR_{IN} = 5 dB; and 5) SNR_{IN} = 0 dB. Thus, a total of ten simulations were used for the training, for one network. In this experiment, the input signals of the NN are the zero to 31st samples (0 to 3.9 kHz) of the FFT amplitude spectrums. The target signals are samples of the FFT amplitude spectrums with a frame corresponding to a training signal for a clean speech signal. Therefore, the target signals for the NN are the zero to 31st samples of the FFT amplitude spectrums.

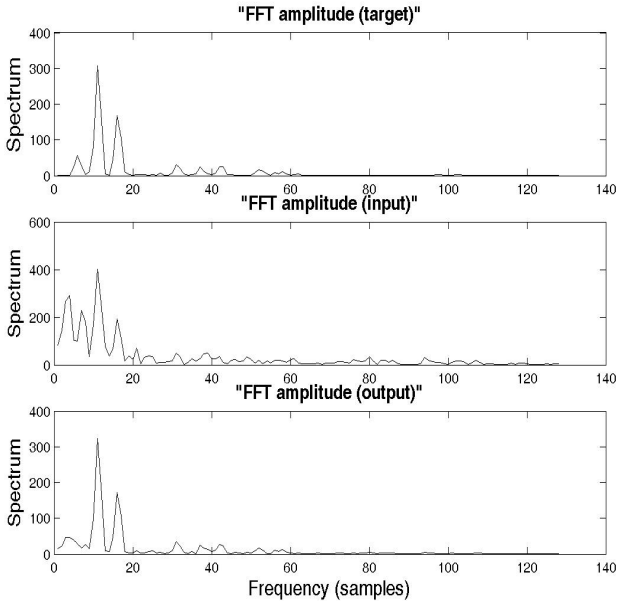

Fig. 4. Comparison of fast Fourier transform (FFT) amplitude spectrums for target, input, and output signal.

In order to highlight the noise elimination performance by the NRNN system, car noise was added to the original clean signal "MAT_32O2A.08", which was spoken by a male speaker, for the tenth frame. Fig. 4 shows examples of the FFT amplitude spectrums for the target (clean speech), input (noisy speech), and output signal, while Fig. 5 shows the FFT phase spectrums for the target, input, and output signal, in the case of SNR_{IN} = 5 dB. Therefore, these figures show that the background noise was significantly reduced when the proposed NRNN system was used.

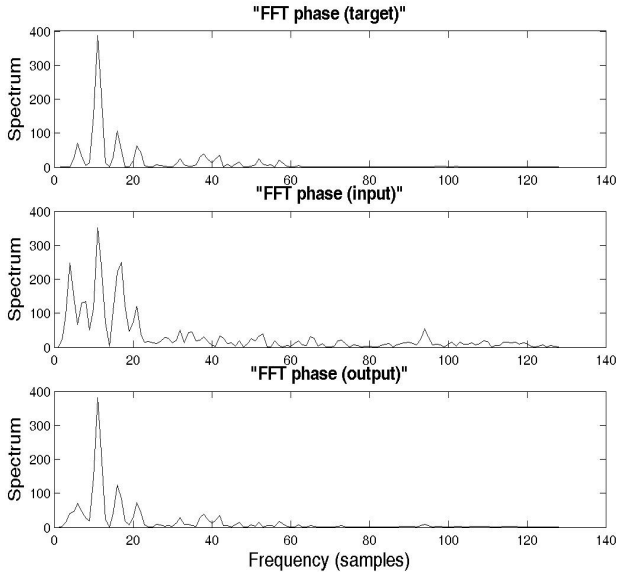

Fig. 5. Comparison of fast Fourier transform (FFT) phase spectrums for target, input, and output signal.

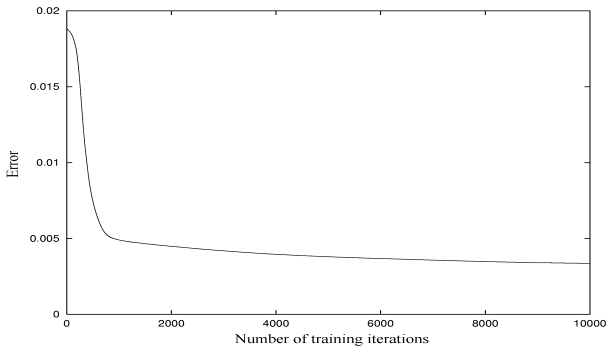

Fig. 6. Error curves of training for neural network with amplitude spectrums.

The training coefficient was set at 0.1 and the inertia coefficient set at 0.03 because these training coefficients produced the best convergence of the NN. The maximum number of training iterations was discontinued after 10,000 times because there was almost no decrease in the training error curves at the minimum error points. The error in Fig. 6 is an average value among ten trials for one network. Fig. 6 shows the training error curves for the NN based on the amplitude spectrums in the case of SNR_{IN} = 5 dB.

III. DESIGN OF PROPOSED SPEECH PROCESSING SYSTEM

In order to consider the practical applications for noise reduction, we adopted the speech processing system as shown in Fig. 7. The proposed system mainly consists of the weighted spectral average filter for reducing unexpected peaks at frames, the NN for training two kinds of different patterns with the FFT amplitude and phase spectrums, and the lateral inhibition function for obtaining a significant auditory spectral representation. In Fig. 7, first, the noisy speech signal *x(t)*, sampled at 8 kHz, is multiplied by a 64 sample Hamming window. After passing through the FFT, the Fourier-transformed signal is separated into its amplitude and phase spectrums. Next, the amplitude spectrums are delayed by 3 frames and averaged using the weighted spectral average filter for each frame. The output signals are added to the input signals for the NRNN system. In the case of the amplitude spectrums, the proposed system detects voiced sections in every frame where $R_f \geq T_h$, and transitional sections in every frame where $R_f < T_h$. Here, R_f is the effective value obtained for each frame and T_h is the threshold value which is determined experimentally. In this experiment, the maximum value for the threshold calculated at the start of seven sections was used [10, 11].

At the output of the decision of the voiced and transitional sections block, there is a switch to send the FFT or log power spectrums in Fig. 7. When the detected results for each frame are voiced sections, we use FFT power spectrums to convolve with the lateral inhibition function, called algorithm I. When the detected results for each frame are transitional sections, we use log power spectrums and a spectrum smoothing method by using the average filter, that is, algorithm II. Some negative values after convolution do not contain any useful information in the present situation, and therefore are set at zero. In this experiment, the coefficients for the lateral inhibition function were "*B* = 6" and "□ = 2.0" as shown in Fig. 1, where "*B*" shows the width of the lateral inhibition function and "□" shows the whole amplitude of the lateral inhibition function. After obtaining the amplitude and phase spectrums, the output of the speech signal is regenerated by an inverse fast Fourier transform (IFFT) [10, 11].

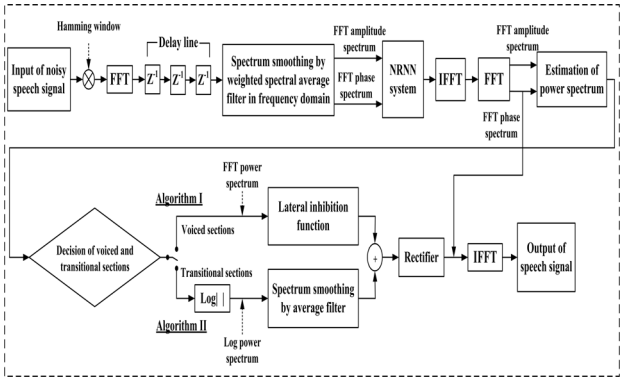

Fig. 7. Proposed speech processing system.

IV. EXPERIMENTAL RESULTS AND CONSIDERATIONS

Using the basic composition conditions described above, experiments confirmed that the proposed system was effective for speech degraded by additive car and street noise based on measuring the SNR. To evaluate the performance, noisy speech data for twenty different test utterances were randomly selected from test sets A, B, and C [10, 11].

The proposed system was compared with a conventional spectral subtraction (SS) method [3] and the minimum mean-square error log-spectral amplitude (MMSE-LSA) [8] estimator method for car and street noise. The SS method is the classic algorithm that is used in speech enhancement and noise suppression. This method estimates the noise power spectrums from the magnitude spectrums of the noisy speech measured during non-speech activity and then subtracts the noise power spectrums from the speech power spectrums for each frame. The enhanced speech is reconstructed, through an inverse Fourier transform, from both the enhanced magnitude spectrums and the original phase spectrums [3]. When implementing the SS method in this experiment, Hamming-windowed frames that overlapped by 50% were used to reduce edge effects. Moreover, the MMSE-LSA estimator method is based on a minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [9], which can be derived by modeling speech and noise spectrums as statistically independent Gaussian random variables. This MMSE-LSA amplitude estimator method is obtained by means of minimizing the mean-squared error of the log-power spectrums, and the gain function derived in Ephraim and Malah [8] is constructed to minimize the mean-square error estimates of log spectrums. When implementing these conventional methods, the frame length was 64 samples (8 ms) and the overlap was 32 samples (4 ms). At each frame, a Hamming window was used [10, 11].

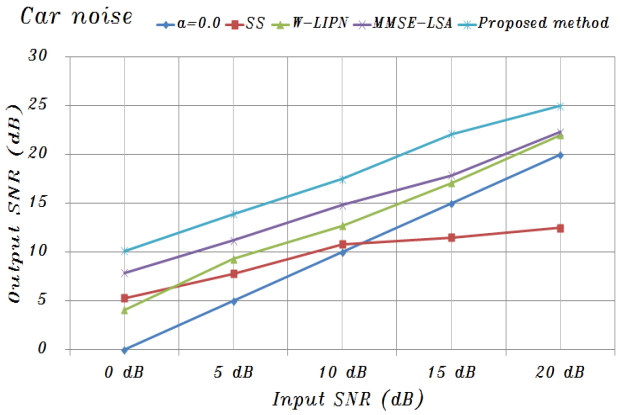

Fig. 8. A comparison of the proposed method and conventional noise suppression methods with the addition of car noise. SNR: signal-to-noise ratio, SS: spectral subtraction, NRNN: noise reduction neural network, W-LIPN: without lateral inhibition processing and NRNN, MMSE-LSA: minimum mean-square error log-spectral amplitude.

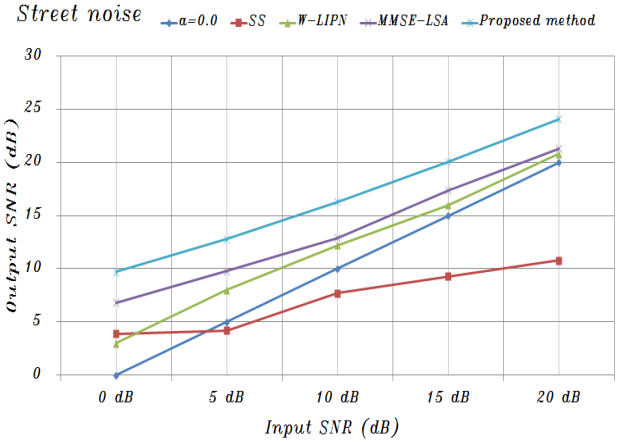

Fig. 9. A comparison of the proposed method and conventional noise suppression methods with the addition of street noise. SNR: signal-to-noise ratio, SS: spectral subtraction, NRNN: noise reduction neural network, W-LIPN: without lateral inhibition processing and NRNN, MMSE-LSA: minimum mean-square error log-spectral amplitude.

Figs. 8 and 9 show the averages of the approximate SNR_{OUT} values over the test utterances for the proposed system, as compared with the SS method and the MMSE-LSA estimator method at different noise levels (SNR_{IN} = 20 to 0 dB) for each noise. In the case of stationary noise, as shown in Fig. 8, the maximal improvement in the SNR_{OUT} values, with the proposed method, was approximately 3.0 dB, 4 dB, and 11.5 dB better for car noise, when compared with the MMSE-LSA, without lateral inhibition processing and NRNN (W-LIPN), and SS methods, respectively, where the W-LIPN shows the maximal improvement in the SNR_{OUT} values obtained by not using the lateral inhibition function or the NRNN system, in the proposed speech processing system of Fig. 7. Moreover, a similar tendency was found for non-stationary noise as shown in Fig. 9, i.e., the maximal improvement in

the SNR_{OUT} values, with the proposed method, was approximately 2.5 dB, 3.5 dB, and 11.0 dB better for street noise, when compared with the MMSE-LSA, W-LIPN, and SS methods. The maximal improvement in the SNR_{OUT} values with the proposed method was approximately 10.0 dB better for car noise and 8.0 dB better for street noise, when compared with = *0.0*. Thus, the 10.0 dB improvement in the SNR_{OUT} was quite significant, and was evident when listening to the output.

V. CONCLUSIONS

A speech processing system was proposed that uses a lateral inhibition mechanism model and noise reduction neural network based on amplitude and phase spectrums to reduce background noise. In summary, the experimental results were as follows: 1) The noise reduction with the function of lateral inhibition was different for car and street noise, and was especially remarkable for car noise; and 2) The noise reduction was significant under serious SNR_{IN} conditions of up to about 0 dB for speech data.

The proposed system using the function of lateral inhibition was experimentally demonstrated as an effective noise suppression system for white, car, restaurant, subway, and street noise. Therefore, it is believed that the present research results will be useful for noise suppression and speech enhancement under noisy conditions.

REFERENCES

[1] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, pp. 4041-4044, 2008.

[2] S. Jeong, H. Yang, and M. Hahn, "Two-channel noise reduction for robust speech recognition in car environments," *Electronics Letters*, vol. 44, no. 17, pp. 1042-1043, 2008.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[4] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 4, pp. 354-358, 1978.

[5] L. Y. Sui, X. W. Zhang, J. J. Huang, and B. Zhou, "An improved spectral subtraction speech enhancement algorithm under non-stationary noise," *Proceedings of International Conference on Wireless Communications and Signal Processing*, Nanjing, China, pp. 1-5, 2011.

[6] K. Daqrouq, I. N. Abu-Isbeih, and M. Alfauri, "Speech signal enhancement using neural network and wavelet transform," *Proceedings of the 6th International Multi-Conference on Systems, Signals and Devices*, Djerba, Tunisia, pp. 1-6, 2009.

[7] W. G. Knecht, M. E. Schenkel, and G. S. Moschytz, "Neural network filters for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 433-438, 1995.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443-445, 1985.

[9] R. Okamoto, Y. Takahashi, H. Saruwatari, and K. Shikano, "MMSE STSA estimator with nonstationary noise estimation based on ICA for high-quality speech enhancement," *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Dallas, TX, pp. 4778-4781, 2010.

[10] J. S. Choi and S. J. Park, "Speech enhancement system based on auditory system and time-delay neural network," *Proceedings of the 8th International Conference on Adaptive and Natural Computing Algorithms*, Warsaw, Poland, pp. 153-160, 2007.

[11] J. S. Choi, "An adaptive speech enhancement system based on noise level estimation and lateral inhibition," *Acta Acustica united with Acustica*, vol. 93, no. 4, pp. 632-644, 2007.

[12] C. Glackin, L. Maguire, and L. McDaid, "Feature extraction from spectro-temporal signals using dynamic synapses, recurrency, and lateral inhibition," *Proceedings of International Joint Conference on Neural Networks*, Barcelona, Spain, pp. 1-6, 2010.

[13] D. Pearce and H. G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, pp. 29-32, 2000.

Jae Seung Choi

received the B. S. degree in Electronics Engineering from Chosun University, Gwangju, Korea, in 1989, and the M. S. and Ph. D. degrees in Information and Communication Engineering from Osaka City University, Osaka, Japan, in 1995 and 1999, respectively. From 2000 to 2001, he was a researcher with AVC Company of Matsushita Electric Industrial Co., Ltd., Osaka, Japan. Since 2002 he has been a project leader with the Digital Technology Research Center of Kyungpook National University. Since 2007, he has been with Silla University where he is currently professor in the Department of Electronic Engineering. His research interests are in the areas of digital signal processing, speech signal processing, and digital communications.