A Spectrum Recovery Algorithm using Signal-to-Noise Ratio Classification for Noise Reduction

Jae Seung Choi

Department of Electronic Engineering, College of Engineering, Silla University, 140 Baegyang-daero (Blvd), 700 Beon-gil (Rd), Sasang-gu, Busan, 617-736, Korea

jschoi@silla.ac.kr

Abstract

In the area of speech signal processing, real background noise is important problem for noise reduction, therefore more skillful methods are required in this area. Accordingly, this paper proposes a spectrum recovery algorithm using a signal-to-noise ratio classification method based on a classification of a voiced or unvoiced signal. Therefore, the proposed algorithm recovers a speech spectrum from a noisy speech spectrum using a time-delay neural network for noise reduction. As such, the proposed system detects the voiced and unvoiced signal, then reduces the noise spectrums for each input frame using the time-delay neural network. Based on measuring correct classification rates and spectrum recovery results, experiments confirm that the proposed algorithm is effective for speech degraded by various noises.

Keywords: *Spectrum recovery, classification method, noise reduction, FFT amplitude, time-delay neural network, background noise*

1. Introduction

In the field of speech processing, the treatment of background noise is still an important problem for speech recognition. This kind of noise cannot be simply eliminated with a Wiener filter [1] *etc*., therefore requires more skillful methods. Noise reduction approaches can be generally considered to reduce noise in a conversation under noisy environment, such as spectral subtraction [2-4], adaptive filter [5, 6], minimum mean-square error estimator [7, 8], fuzzy logic [9, 10], neural network [11-14], and time-delay neural network [15-18]. For instance, in a study by Boll [2], the spectral subtraction is effective method for noise reduction. This method attempts to reduce a noise signal, and enhance a speech signal, from a noisy speech signal. However, the enhanced speech signal, by the spectral subtraction method, still shows some residual noise called musical noise, which occurs because of noise-amplitude estimation errors, which are annoying to the human ear. Accordingly, to solve the above-mentioned problem, this paper proposes a spectrum recovery algorithm using a signal-to-noise ratio (SNR) classification method and a time-delay neural network (TDNN) [15-18] based on a classification of a voiced or unvoiced signal, in various noisy environments.

In many speech processing applications for noise reduction and speech enhancement, it is necessary to recover a signal from an amplitude or phase spectrum obtained by a Fourier transform, in order to improve the performance of a system in noisy environments. Several

algorithms for the recovery of the signal from its amplitude or phase spectrum have been proposed [19, 20]. For instance, in a study by Yegnanarayana *et al.*, [19], a signal recovery method is used to recover such signal from the phase spectrum of a short-time Fourier transform. A fundamental method used to recover the signal from the phase spectrum of the short-time Fourier transform is a consecutive extrapolation. The first section is recovered using an iterative algorithm based on phase information of the short-time Fourier transform to accurately recover this segment of a signal sample. The other sections are recovered by using an overlapped rectangular-window at each frame. Futoshi Asano *et al*., [20] used a recovering method, which recovers an LPC (Linear Predictive Coding) spectrum from a microphone-array input speech disturbed by less-directional ambient noise using a coherent subspace method. The speech and noise are classified by using the coherent subspace method in the subspace domain. However, the classification by the coherent subspace method is only effective in the case of high SNR conditions.

Many approaches previously exit that suppress noise in a conversation under noisy condition, such as neural filters based on a nonlinear adaptive filter for noise cancellation [21-27]. Neural network and TDNN have been reported in applications such as pattern recognition, speech recognition, noise reduction, and adaptive beam-former. Tamura and Waibel [21, 22] introduced a noise reduction algorithm using a neural network to map from a noisy speech signal to a clean speech signal. In this study, a four-layered feed-forward neural network is used to reduce noise. To discriminate segments of a speech as a voiced or unvoiced signal, a neural network training algorithm, which uses a quasi-Newton error minimization method, is introduced in a study [23]. This discrimination method is based on features computed for each speech segment and used as input to the neural network. The neural network is trained to accomplish a three-class discrimination using input features extracted from each speech frame. The weights of the neural network are trained using a fast training algorithm based on the quasi-Newton error minimization method with a positive-definite approximation of a Hessian matrix. Yoganathan and Moir [24] introduced an adaptive time delay neural network using a nonlinear noise canceller to suppress a noise signal from a noisy speech signal. This study presents a nonlinear switched Griffiths-Jim beamformer structure using a fragmentally connected three-layer TDNN and adaptive noise canceller (ANC). The construction of the TDNN is composed of a three-layer feed-forward neural network. The TDNN is trained using an error back propagation learning algorithm.

On the other hand, in our previous work [18], a neural network needs to be constructed using a time structure, as the time variation is significant information. Moreover, an amplitude component contains more information than a phase component when a speech signal is generated by a fast Fourier transform (FFT). Accordingly, this paper proposes an algorithm that restores the FFT amplitude component using the TDNN [15-18], which includes a time structure in the neural network as a method of spectrum recovery, then confirms the efficiency of the proposed algorithm based on experiments of noise reduction in a speech signal. Using the correct classification rates and spectrum recovery results, experiments confirm that the proposed algorithm is effective for speech degraded by noises, such as white and subway noise.

The remainder of this paper is organized as follows. Section 2 describes a noisy speech signal to use in this study. Section 3 introduces the construction of the proposed time-delay neural network. Section 4 explains a speech and noise database used in the experiments and discusses the experimental results when using the proposed algorithm. Section 5 presents some final conclusions.

2. Noisy Speech Signal

The noisy speech signal is assumed to be $s(k)$, and the speech signal disturbed by noise is given by

$$x(k) = s(k) + n(k) \quad (1)$$

The fast Fourier transform (FFT) for equation (1) is given by equation (2).

$$X(\omega) = S(\omega) + N(\omega) \quad (2)$$

Where

$$x(k) \leftrightarrow X(\omega),$$

$$X(\omega) = \sum_{k=0}^{L-1} x(k) e^{j\omega k},$$

$$x(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) e^{j\omega k} d\omega$$

Here, $n(k)$ is white and subway noise with a sampling frequency of 8 kHz, where white noise was generated by a computer program with a sampling frequency of 8 kHz. In addition, the subway noise included in the Aurora-2 database was also used in this experiment.

3. Proposed Time-Delay Neural Network (TDNN)

This paper proposes an algorithm that restores the FFT amplitude component using the TDNN as a method of spectrum recovery, because the FFT amplitude component contains more speech information than a phase component. Figure 1 shows the construction of the proposed TDNN system. First, a noisy speech signal $x(k)$ is divided into length frames of 128 samples (16 ms). Next, the $x(k)$ is detected in the voiced or unvoiced sections, then separated into FFT amplitude components according to the voiced or unvoiced sections. Thereafter, the separated FFT amplitude components are added to the appropriate TDNNs with low, mid and high frequency bands.

In this experiment, the TDNNs for the low, mid, and high frequency band are composed of four layers and the compositions of the TDNNs are 22-60-22-22. Input signals for the TDNNs with the low and mid frequency bands are 0 to 21st samples and 22nd to 43rd samples of the FFT amplitude component, respectively, where the input signals consist of a target frame, two previous frames, and the following frame. Target signals for the TDNNs with the low and mid frequency bands are 0 to 21st samples and 22nd to 43rd samples of the FFT amplitude component with a frame corresponding to a training signal for a clean speech signal, respectively. Meanwhile, the input signals for the TDNN with the high frequency band are 43rd to 64th samples of the FFT amplitude component, where the input signals also contain additional frames (the target frame, the two previous frames, and the following frame). The target signals are 43rd to 64th samples of the FFT amplitude component with a frame corresponding to the training signal for the clean speech signal.

A final FFT amplitude component is then obtained by combining the results from the TDNNs with the low, mid and high frequency bands. However, the FFT phase component is directly obtained from an original noisy speech signal, after detecting voiced and unvoiced sections. Thereafter, an enhanced speech signal $y(k)$ is regenerated using an inverse fast Fourier transform (IFFT). In normalization process, the $x(k)$ was normalized by an effective

value e_m, then adjusted by the level of a constant value for the level of the noisy speech signal of SNR_{in} = 10 dB as the normal standard. Therefore, the effective values for all utterances were adjusted to this level. Here, e_m is the level of the effective value obtained for each entire utterance and it is represented as follow equation.

$$e_m = \sqrt{\sum_{m=1}^{M} x^2(m) / M} \qquad (3)$$

Where *M* is the number of samples in each utterance.

Table 1 shows the parameters used to implement the training and other conditions for each TDNN used in this experiment. In the training of the proposed TDNNs, the training coefficient was set to 0.2 and the inertia coefficient was set to 0.7. Moreover, random numbers from –1.0 to 1.0 is used as an initial weight, and 10,000 was set as the maximum number of training iteration for the experiment. In this experiment, the proposed TDNNs are trained using a back propagation algorithm [28, 29].

Table 1. Various Conditions for Training of TDNN

Initial weight	Radom numbers from -1.0 to 1.0
Coefficient of training	α = 0.2
Coefficient of inertia	β = 0.7
Maximum training iteration	10,000 times
Construction of network	22-60-22-22

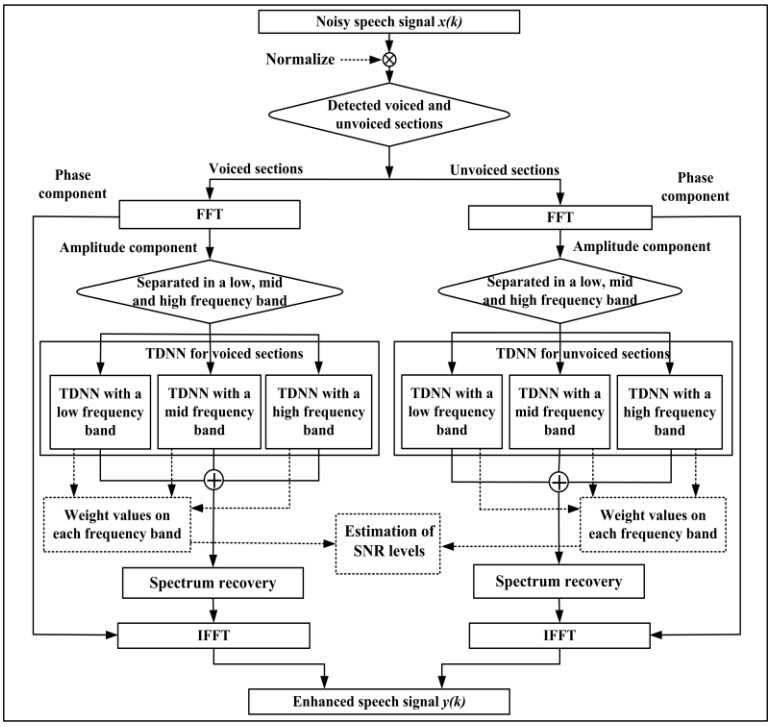

Figure 1. The Construction of the Proposed TDNN System

4. Experimental Results

In this section, experiment results confirmed that the proposed algorithm was effective for speech degraded by white and subway noise based on measuring the classification rates and spectrum recovery results, using the basic composition conditions described above.

4.1. Speech and Noise Database

To test the performance of the proposed TDNN system, the speech and noise data used in this experiment is presented in this section.

All speech data used in this experiment was the Aurora-2 database that consists of English connected digits recorded in clean environments with a sampling frequency of 8 kHz [30]. The speech data of the Aurora-2 database is distributed by ETSI (European Telecommunications Standards Institute) committee and is derived from a subset of the TI-Digits database [31], which consists of English-connected digits spoken by American English speakers. Eight different background noises have been added to the speech data at different signal to noise ratios (SNRs). The speech data is down sampled from 20 kHz to 8 kHz with a low-pass filter and filtered with a G712 characteristic [32]. These speech data are considered as clean speech data. These clean speech data are artificially contaminated by adding eight different types of real-life background noises (subway (inside a subway), babble (crowd of people), car, exhibition hall, restaurant, street, airport, and train station noises) to the clean speech data at several SNR levels (20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB, clean (no noise added)), where street and babble noises are non-stationary and other noises are stationary. Because the main part of the energy in speech signals is concentrated in lower frequency areas and spectrums of these noises looks extremely like the spectrums of speech signal data, it is thought that the classification of background noise from speech signal data is not easy.

The Aurora-2 database offers two different training modes: (1) clean training mode, *i.e*. training on clean speech data only, (2) multi-conditional training mode, i.e. training on clean speech and noisy speech data. The clean training mode includes 8440 clean utterances selected from the training part of the English-connected digits; which contains the voices of 55 male and 55 female adult recordings. The same 8440 speech data are also used in the multi-conditional training mode.

4.2. Classification Tests by Proposed TDNN System

In this experiment, the proposed algorithm was evaluated using speech data from the Aurora-2 database in Test Sets A, B, and C and two types of background noise, *i.e.*, subway noise in Test Set A, and white noise generated by a computer program. In the experiments, the total time duration of the noise data was about 23 second for white and subway noise. In this experiment, the TDNNs are trained using noisy speech data artificially added at several SNRs (20 dB, 15 dB, 10 dB, 5 dB, and 0 dB). When using the Aurora-2 database, the TDNNs are trained after adding white and subway noise to the clean speech data in the Aurora-2 database.

The performance of the proposed TDNN system was tested based on the correct classification rate, frame-by-frame, and the definition of the classification rate was the ratio of the number of frames in which the SNR levels were correctly estimated to the total number of frames given as the input. In this experiment, the total number of frames was about 100 to 300 when included silent frames, which were included as the proportion about 15% for short utterances and about 20% for long utterances.

Figure 2 and Figure 3 show the average values of the classification rates of the proposed TDNN system for the noises, when using a total of twenty different test utterances selected

from Test Sets A and B. In the case of TDNN with the low frequency band when voiced sections, the classification rates averaged over fifty utterances were 90% or more for each condition of white and subway noises in Test Sets A and B. Moreover, the classification rates were 87% or more for each condition of white and subway noises, in the case of TDNN with the low frequency band when unvoiced sections. However, the average values of the classification rates were approximately 2.5% worse for such noises, in the case of TDNN with the mid and high frequency bands when voiced and unvoiced sections, respectively.

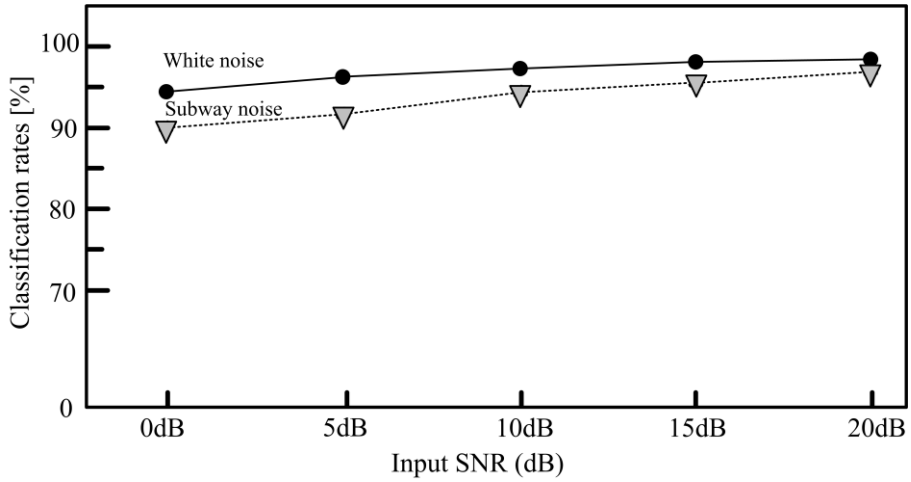

Figure 2. Classification Rates for TDNN with the Low Frequency Band when Voiced Sections

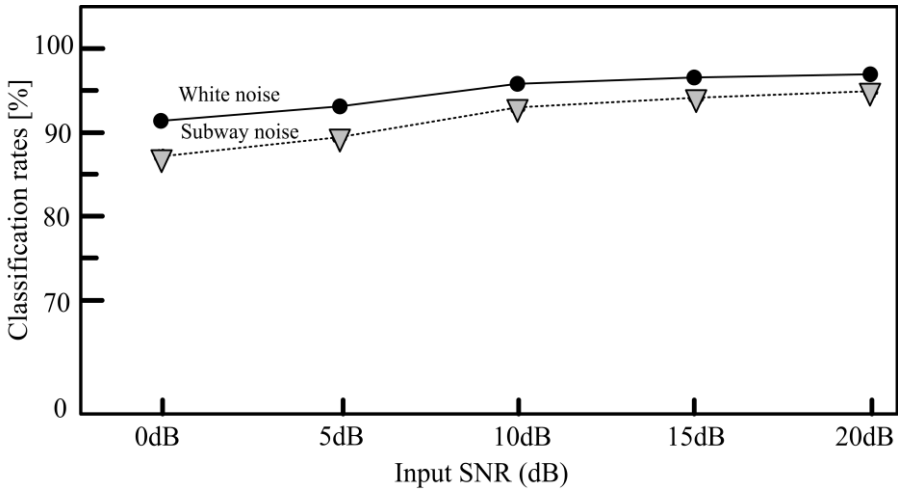

Figure 3. Classification Rates for TDNN with the Low Frequency Band when Unvoiced Sections

4.3. Experimental Results of Spectrum Recovery

Since the purpose of this paper is to reduce the noise in the noisy speech signal using the TDNN system, the experimental results for noise reduction are described for speech data. In this experiment, the effectiveness of the proposed TDNN system was confirmed with the proposed algorithm under conditions up to about SNR_{in} (input SNR) = 0 dB when using the

SNR. To investigate the general property, the effectiveness for noise reduction was evaluated using trained and non-trained speech data when white and subway noise were added to a clean speech signal. Figure 4, Figure 5 and Figure 6 show a comparison of the FFT amplitude component based on training the TDNN system for one frame selected from Aurora-2 speech signal "93528" spoken by a female speaker, when white noise was added to the clean speech signal in the case of $SNR_{in} = 0$ dB, respectively. Figure 4 shows a comparison of the FFT amplitude component for the target signal (solid line) and input signal (dotted line) when adding white noise. Figure 5 shows a comparison of the FFT amplitude component for the target signal (solid line) and output signal (dotted line) when using trained speech data and adding white noise. Figure 6 shows a comparison of the FFT amplitude component for the target signal and output signal when using non-trained speech data and adding white noise. In particular, the proposed TDNN system restored the FFT amplitude components for white noise, and reduced white noise by concentrating on the low frequency band. Accordingly, these figures show that background noise was significantly reduced when using the proposed TDNN system.

Accordingly, from the above-mentioned results, the possibility of effective noise reduction using the proposed TDNN system was confirmed for both a trained noisy speech signal and a non-trained noisy speech signal, regardless of the kind of noise. In fact, the noise reduction was remarkable under conditions up to about SNR_{in} = 0 dB, especially for white noise.

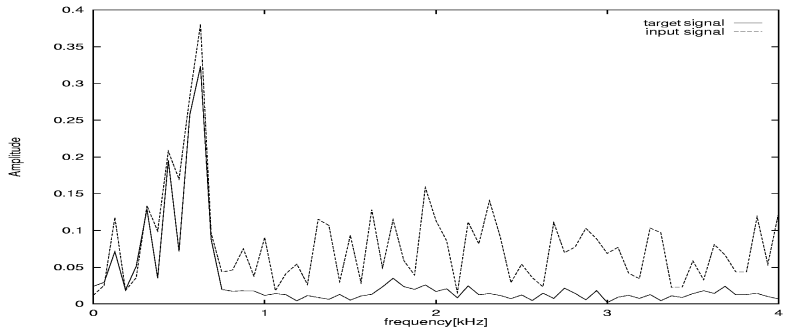

Figure 4. Comparison of FFT Amplitude Component for Target Signal and Input Signal when Adding White Noise

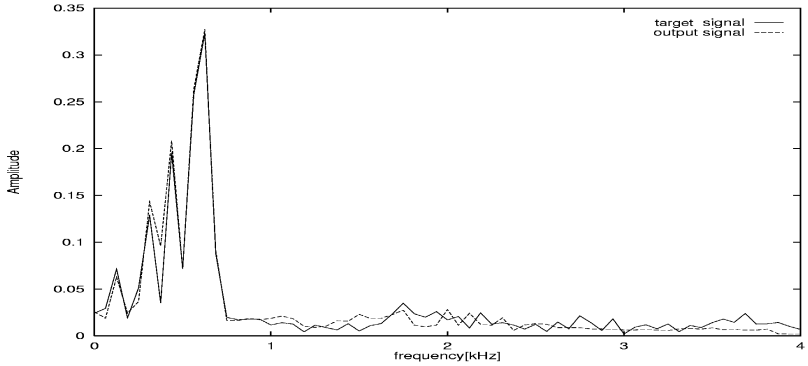

Figure 5. Comparison of FFT Amplitude Component for Target Signal and Output Signal when using trained Speech Data and Adding White Noise

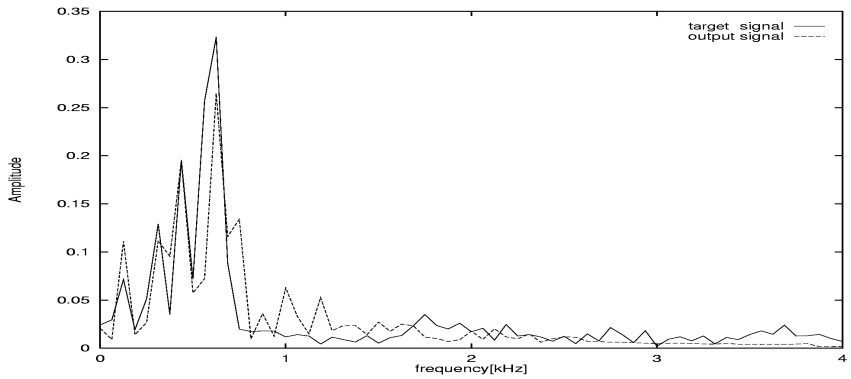

Figure 6. Comparison of FFT Amplitude Component for Target Signal and Output Signal when using Non-trained Speech Data and Adding White Noise

5. Conclusions

A TDNN system based on classification of a voiced or unvoiced signal was proposed that uses a TDNN to reduce background noise. Experimental results confirmed that the proposed algorithm is effective for white and subway noise, as demonstrated by the classification rates and spectrum recovery results. In summary, the experimental results were as follows:

1. The possibility of noise classification using a TDNN was confirmed in this experiment.

2. The noise reduction was significant under input SNR conditions of up to about 0 dB for sentences.

3. Background noise was significantly reduced when using the proposed TDNN system.

4. The effect of noise reduction was significant for white and subway noise, and especially remarkable for white noise.

The following problems remain as future areas for study.

1. The effectiveness of the proposed algorithm needs to be evaluated for speech degraded by heavy noise and various non-stationary noises in a real environment.

As mentioned above, the proposed algorithm using the TDNN was experimentally demonstrated for white and subway noise. Therefore, it is believed that the present research results will be useful for the speech recognition under noisy conditions.

References

[1] T. V. Sreenivas and P. Kirnapure, "Codebook constrained wiener filtering for speech enhancement", IEEE Transactions on Speech and Audio Processing, vol. 4, no. 5, **(1996)**, pp. 383-389.
[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech, Signal Processing, vol. 27, no. 2, **(1979)**, pp. 113-120.
[3] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise", IEEE Trans. Acoust., Speech, Signal Processing, vol. 6, no. 5, **(1978)**, pp. 471-472.
[4] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence", IEEE Trans. Signal Processing, vol. 39, no. 9, **(1991)**, pp. 1943-1954.
[5] M. R. Sambur, "Adaptive noise cancelling for speech signals", IEEE Trans. Acoust., Speech, Signal Processing, vol. 26, no. 5, **(1978)**, pp. 419-423.
[6] J. W. Kim and C. K. Un, "Enhancement of noisy speech by forward/backward adaptive digital filtering", IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 86, no. 11, **(1986)**, pp. 89-92.

[7] 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, **(1985)**, pp. 443-445.
[8] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, **(1984)**, pp. 1109-1121.
[9] D. O. Aborisade, "A Novel Fuzzy logic Based Impulse Noise Filtering Technique", International Journal of Advanced Science and Technology (IJAST), vol. 32, **(2011)** July, pp. 79-88.
[10] D. O. Aborisade, "Novel Fuzzy logic Based Edge Detection Technique", International Journal of Advanced Science and Technology (IJAST), vol. 29, **(2011)** April, pp. 75-82.
[11] 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, **(2009)**, pp. 1-6.
[12] W. G. Knecht, M. E. Schenkel and G. S. Moschytz, "Neural network filters for speech enhancement", IEEE Trans. Speech and Audio Processing, vol. 3, no. 6, **(1995)**, pp. 433-438.
[13] J. Dheeba and J. G. Wiselin, "Detection of Microcalcification Clusters in Mammograms using Neural Network", International Journal of Advanced Science and Technology (IJAST), vol. 19, **(2010)** June, pp. 13-22.
[14] M. V. Ishwarya, "An Improved Online Tamil Character Recognition Using Neural Networks", International Journal of Advanced Science and Technology (IJAST), vol. 42, **(2012)** May, pp. 1-10.
[15] M. Debyeche, A. Amrouche and J. P. Haton, "Distributed TDNN-Fuzzy Vector Quantization For HMM Speech Recognition", International Conference on Multimedia Computing and Systems, **(2009)** April, pp. 72-76.
[16] R. A. Mitchell and A. Shaw, "Vowel recognition with a time-delay neural network", IEEE International Conference on Systems Engineering, **(1990)**, pp. 637-640.
[17] J. B. Hampshire and A. H. Waibel, "A novel objective function for improved phoneme recognition using time delay neural networks", IEEE Transactions on Neural Networks, vol. 1, no. 2, **(1990)**, pp. 216-228.
[18] J. S. Choi and S. J. Park, "Speech Enhancement System based on Auditory System and Time-Delay Neural Network", 8th International Conference on Lecture Notes in Computer Science, Part II, **(2007)** April, pp. 153-160.
[19] B. Yegnanarayana, S. Fathima and H. Murthy, "Reconstruction from Fourier transform phase with applications to speech analysis", IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 12, **(1987)** April, pp. 301-304.
[20] F. Asano and S. Hayamizu, "Speech enhancement using CSS-based array processing", 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, **(1997)** April, pp. 1191-1194.
[21] S. Tamura and A. Waibel, "Noise reduction using connectionist models", 1988 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-88), vol. 1, **(1988)** April, pp. 553-556.
[22] S. Tamura, "An analysis of a noise reduction neural network", 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89), vol. 3, **(1989)** May, pp. 2001-2004.
[23] T. Ghiselli-Crippa, "A. El-Jaroudi, Voiced-unvoiced-silence classification of speech using neural nets", IJCNN-91-Seattle International Joint Conference on Neural Networks, vol. 2, **(1991)** July, pp. 851-856.
[24] V. Yoganathan and T. J. Moir, "Speech enhancement using a nonlinear neural switched Griffiths-Jim beamformer", Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA), **(2010)** May, pp. 217-220.
[25] F. Liew Ban, A. Hussain and S. A. Samad, "Speech enhancement by noise cancellation using neural network", in Proceedings of TENCON 2000, vol. 1, **(2000)**, pp. 39-42.
[26] M. Stella, D. Begusic and M. Russo, "Adaptive Noise Cancellation Based on Neural Network", 2006 International Conference on Software in Telecommunications and Computer Networks (SoftCOM 2006), **(2006)** September-October, pp. 306-309.
[27] F. Li and G. Xu, "Quantum BP Neural Network for speech enhancement", Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications, vol. 2, **(2009)** November, pp. 389-392.
[28] S. Wang, X. Ling, F. Zhang and J. Tong, "Speech Emotion Recognition Based on Principal Component Analysis and Back Propagation Neural Network", International Conference on Measuring Technology and Mechatronics Automation, vol. 3, **(2010)** March, pp. 437-440.
[29] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", Proc. ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, **(2000)**.
[30] R. G. Leonard, "A database for speaker independent digit recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, **(1984)** March, pp. 328-331.

[31] ITU-T (International Telecommunication Union) recommendation G. 712, Transmission performance characteristics of pulse code modulation channels, **(1996)**, pp. 1-31.

Author

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 Digital Technology Research Center of Kyungpook National University. Since 2007, he has been with the Silla University where he is currently an associate professor in the Department of Electronic Engineering. His research interests are in the areas of speech signal processing, speech recognition, adaptive signal processing, noise reduction, and neural network.