Chapter 29
Time-Delay Neural Network with 3 Frequency Bands Based on Voiced Speech Discrimination in Noise

Jae Seung Choi

Abstract Information on the time variation in a speech signal is significant when training a neural network for the speech signal input. Therefore, this paper proposes a time-delay neural network with 3 frequency bands based on voiced speech discrimination in the condition of background noises. The effectiveness of the proposed network is experimentally confirmed based on measuring the correct discrimination rates for speech degraded by various noises.

Keywords Discrimination rate · Voiced speech discrimination · Time-delay neural network · Frequency band · Background noise

29.1 Introduction

Recent developments in speech recognition have resulted in various studies, such as neural network [1–4], time-delay neural network [5, 6], hidden Markov model (HMM) [3], and so on.

In the area of speech signal processing, the major application of a neural network (NN) is the category classification of phoneme recognition, while in the area of speech enhancement and noise reduction, the major application of the NN is the extraction of speech sections from a noisy speech signal [7, 8]. Thus, for speech recognition, the NN 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 a time-delay neural network (TDNN) [5, 6] system with 3 frequency bands based on

J. S. Choi (✉)
Department of Electronic Engineering, College of Engineering, Silla University,
140 Baegyang-daero (Blvd), 700 Beon-gil (Rd), Sasang-gu, Busan 617-736, Korea
e-mail: jschoi@silla.ac.kr

H.-K. Jung et al. (eds.), *Future Information Communication Technology and Applications*,
Lecture Notes in Electrical Engineering 235, DOI: 10.1007/978-94-007-6516-0_29,

voiced speech discrimination in the condition of background noises, which includes a time structure in the NN.

To evaluate the proposed TDNN system, the performance of the proposed system is evaluated based on correct discrimination rates at frame-by-frame for white, car, restaurant, and subway noise.

The remainder of this paper is organized as follows. Section 29.2 describes an additive noise model, and Sect. 29.3 introduces the construction of the proposed time-delay neural network. Section 29.4 discusses experimental results when using the proposed system. Section 29.5 presents some final conclusions.

29.2 Additive Noise Model

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

$$x(k) = s(k) + n(k). \tag{29.1}$$

The fast Fourier transform for Eq. (29.1) is given by Eq. (29.2).

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega}) \tag{29.2}$$

where

$$x(k) \leftrightarrow X\left(e^{j\omega}\right), \quad X\left(e^{j\omega}\right) = \sum_{k=0}^{L-1} x(k)e^{j\omega k}, \quad x(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(e^{j\omega}\right)e^{j\omega k}d\omega.$$

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

29.3 Construction of Time-Delay Neural Network (TDNN)

This section describes the construction of proposed TDNNs, which are composed of a low, mid, and high frequency band.

Information on the time variation in a speech signal is significant when training the NN for the speech signal input. Therefore, this paper proposes the three kinds of TDNNs to be constructed the low, mid, and high frequency bands. Figure 29.1 shows the construction of the proposed TDNNs for the low, mid, and high frequency bands used in this experiment. A time series of 32-unit FFT amplitude components is fed into the input layer with n frames. Thereafter, the four frames in the input layer are connected to the frame in the first hidden layer. Every 6 frames in the first hidden

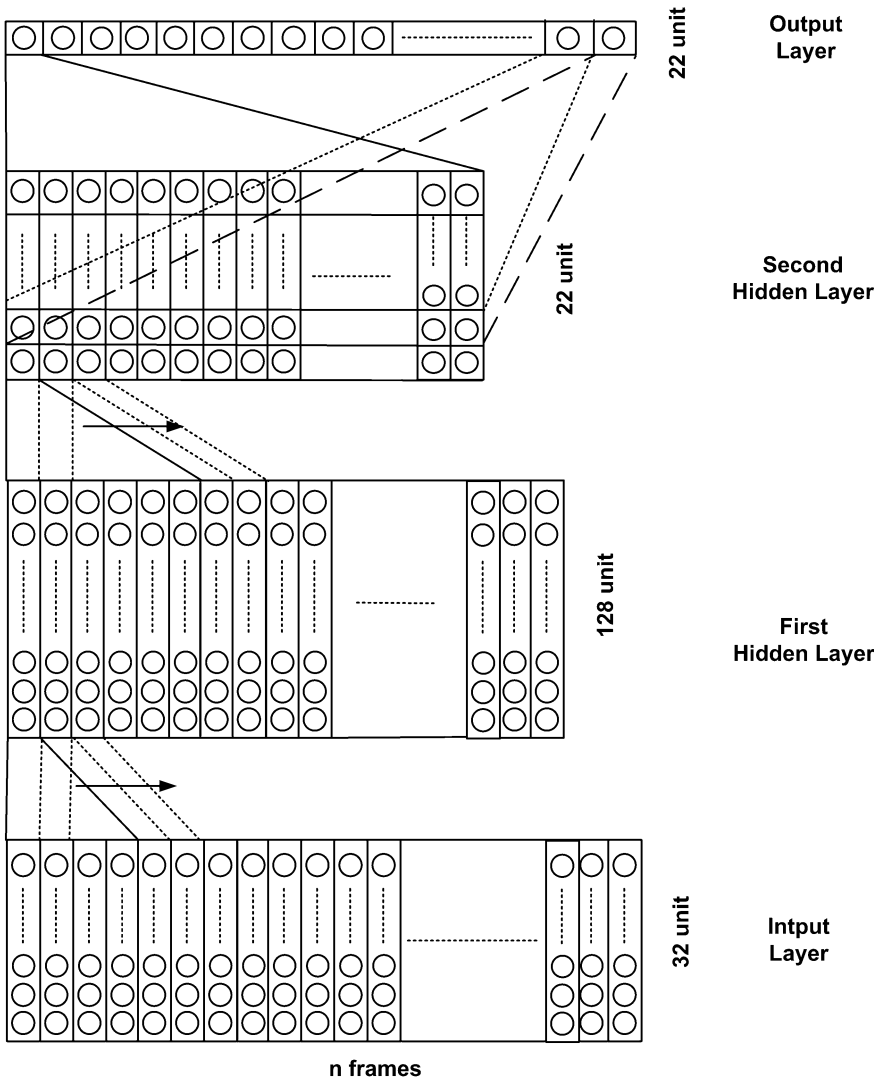

Fig. 29.1 Construction of the proposed TDNNs for low, mid, and high frequency bands

layer with 128 units are connected to the frame in the second hidden layer. Then every frame in the second hidden layer with 22 units is connected to the output layer. Accordingly, TDNNs for the low, mid, and high frequency bands are composed of four layers and the composition of the TDNN is 32-128-22-22. In this experiment, the input signals for the TDNNs with the low, mid, and high frequency bands are the 0 to 31st samples of the FFT amplitude component, respectively, where the input signals consist of the target frame, the two previous frames, and the following frame. The target signals for the TDNNs with the low, mid, and high frequency bands are

the 0 to 31st samples of the FFT amplitude component with a frame corresponding to a training signal for a clean speech signal, respectively. In this experiment, ten simulations for one network were performed to train the proposed TDNNs as the following input signal-to-noise ratio (SNR_{in}) conditions: (1) 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, respectively. Thus, a total of fifty simulations using the same network were performed for each SNR_{in} conditions.

Figure 29.2 shows a schematic diagram using the proposed TDNN system that is divided into voiced sections, making the TDNN easier to train according to a somewhat similar pattern. The proposed TDNN system is also constructed for a low, mid, and high frequency band, allowing more effective correlation of the added information.

In Fig. 29.2, a noisy speech signal $x(k)$ is first detected in the voiced sections, and divided into FFT amplitude components with the low, mid, and high frequency bands. Thereafter, the divided FFT amplitude components are added to the voiced sections of the TDNN, as appropriate. In Fig. 29.2, the FFT amplitude components obtained from the noisy speech signal $x(k)$ are added to the input signals of the TDNNs, while the FFT amplitude components obtained from a clean speech signal $s(k)$ are added to the target signals of the TDNNs. Thereafter, the TDNNs are

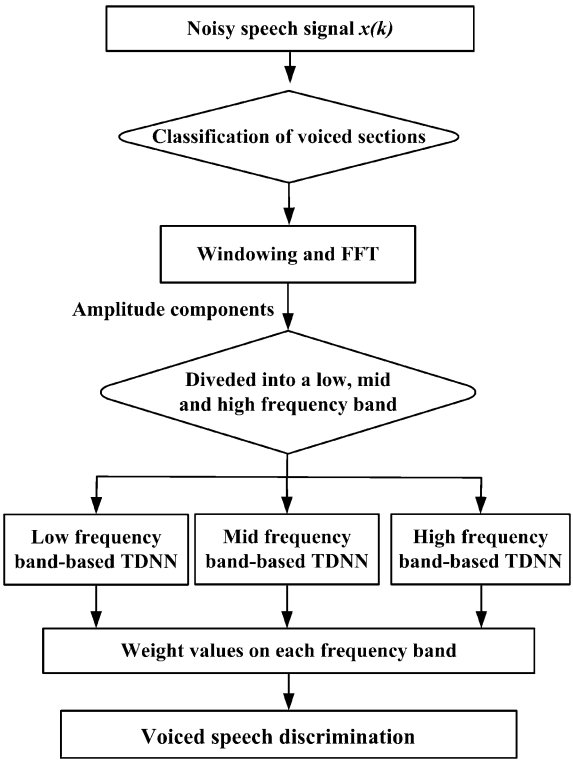

Fig. 29.2 Schematic diagram of the proposed TDNN system

Table 29.1 Various conditions for training of TDNNs

Initial weight	Random numbers from –1.0 to 1.0
Coefficient of training	$\alpha = 0.2$
Coefficient of inertia	$\beta = 0.6$
Maximum training iteration	10,000 times

trained using every individual frame (256 samples). However, for every frame, the two previous frames and following frame are also added to the input signals of the TDNNs, as such frames provide additional information and correspond to the frame of the target signal. Therefore, the total number of input frames is four.

In this experiment, since the amplitude components obtained by the FFT have symmetrical values centered around the 128th component of the FFT, the samples from 0 to the 128th, except for the redundancy area, are divided between the TDNNs with the low, mid, and high frequency bands, and the results added to the TDNNs with the low, mid, and high frequency bands.

Table 29.1 shows the parameters used to implement the training and other conditions for each TDNN used in this experiment. When the training iterations exceeded 10,000, there was almost no decrease in the training error curves at the minimum error points. Therefore, 10,000 was set as the maximum number of training iteration for the experiment. The reason for using a TDNN in this research was to enable car, restaurant, and subway noise with an inconstant intensity in the time domain to be estimated and provide an efficient training algorithm called the back propagation method [9].

29.4 Experimental Results and Considerations

Using the basic composition conditions described above, experiments confirmed that the proposed system was effective for speech degraded by white, car, restaurant, and subway noise based on measuring the discrimination rates.

29.4.1 Speech and Noise Database

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

The 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 [10]. All speech data of the Aurora-2 database is distributed by ETSI committee and is derived from a subset of the TI-Digits database [11], 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 to 8 kHz with a low-pass filter and filtered with a G712 characteristic [12]. 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, 15, 10, 5, 0, −5 dB, clean (no noise added)), where street and babble noises are non-stationary and other noises are stationary. Since the major part of the energy in speech signals is concentrated in the lower frequency areas and the spectra of these noises looks extremely like the spectra of speech signal data, it is thought that the discrimination 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.

In this experiment, the proposed system was tested using speech data from the Aurora-2 database in Test Sets A, B, and C and four types of background noise, i.e. car, subway noise in Test Set A, restaurant noise in Test Set B, and white noise generated by a computer program. Therefore, the proposed TDNNs are trained using noisy speech data artificially added at several SNRs (20, 15, 10, 5, and 0 dB). When using the Aurora-2 database, the TDNNs are trained after adding white, car, restaurant, and subway noise to the clean speech data in the Aurora-2 database. In the experiments, the total time duration of the noise data was about 23 s for white, car, and subway noise, 300 s for restaurant noise, respectively.

29.4.2 Discrimination Test by Proposed TDNN

The performance of the proposed system was tested based on the correct discrimination rate, frame-by-frame, and the definition of the discrimination 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 100–300 when included silent frames, which were included as the proportion about 15 % for short utterances and about 20 % for long utterances.

Figure 29.3 shows the discrimination rates averaged over twenty utterances in the case of TDNN with the low frequency band when voiced sections, for each condition of white, car, restaurant, and subway noise in Test Sets A and B. When the training speech data and the testing data were the same, the average values of

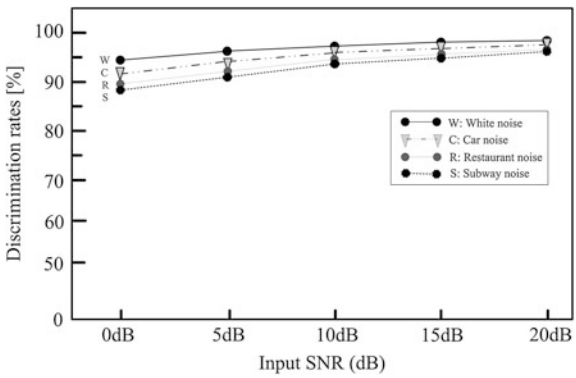

Fig. 29.3 Correct discrimination rates for TDNN with the low frequency band when voiced sections

the discrimination rates were 94 % or more for such noises. However, the average values of the discrimination rates were approximately 4 % worse for such noises, in the case of TDNNs with the mid and high frequency bands when voiced sections, respectively.

29.5 Conclusions

A time-delay neural network with 3 frequency bands based on voiced speech discrimination was proposed for various noises in case of the voiced sections. Experimental results confirmed that the proposed system is effective for white, car, restaurant, and subway noise, as demonstrated by the correct discrimination rates. In the experiment, the discrimination for the voiced speech signal using the proposed TDNNs with 3 frequency bands was confirmed for the input SNR levels. The performance of the proposed system was tested based on the correct discrimination rate. When the training speech data and the testing data were the same, the average values of the discrimination rates were 94 % or more for such noises. However, the effectiveness of the proposed system needs to be evaluated for non-training speech data and speech degraded by heavy noise, in the future.

References

1. Juang CF, Chiou CT, Lai CL (2007) Hierarchical singleton-type recurrent neural fuzzy networks for noisy speech recognition. IEEE Trans Neural Netw 18(3):833–843
2. Knecht WG, Schenkel ME, Moschytz GS (1995) Neural network filters for speech enhancement. IEEE Trans. Speech Audio Process 3(6):433–438
3. Cong L, Asghar S, Cong B (2000) Robust speech recognition using neural networks and hidden Markov models. In: Proceedings of the international on Information technology: coding and computing, pp 350–354

4. Choi JS (2012) Speech processing system using a noise reduction neural network based on FFT spectrums. J Inf Commun Convergence Eng 10(2):162–167
5. Hampshire JB, Waibel AH (1990) A novel objective function for improved phoneme recognition using time delay neural networks. IEEE Trans Neural Netw 1(2):216–228
6. Choi JS, Park SJ (2007) Speech enhancement system based on auditory system and time-delay neural network. In: 8th international conference on lecture notes in computer science. LNCS, Part II, pp 153–160
7. Peng Y, Xiong H, Guo C, Liu H, Zou J (2010) Research on the algorithm of communication network speech enhancement based on BP neural network. Int Conf Adv Comput Theor Eng 3:V3-559–V3-562
8. Vieira K, Wilamowski B, Kubichek R (1997) Speaker verification for security systems using artificial neural networks. Int Conf Ind Electron Control Instrum 3:1102–1107
9. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagation errors. Nature 323:533–536
10. Hirsch H, Pearce D (2000) The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In: Proceedings of the ISCA ITRW ASR2000 on automatic speech recognition: challenges for the next millennium, Paris, France
11. Leonard RG (1984) A database for speaker independent digit recognition. In: IEEE international conference on acoustics, speech, and signal processing, pp 328–331
12. ITU-T (International Telecommunication Union) Recommendation G. 712 (1996) Transmission performance characteristics of pulse code modulation channels, pp 1–31