Mem. Fac. Eng., Osaka City Univ., Vol. 39, pp. 89－93 (1998)

Discriminating method of noise quantity in degraded speech and application for speech enhancement

Jae-seung CHOI [1], Shigeyoshi NAKAJIMA [2], Jiro OKAMOTO [3], Shoichi HOSOKAWA [4] and Yutaka SUZUKI [5]

(Received September 30, 1998)

Synopsis

In order to reduce the noise in the noisy speech, it is desirable to change the parameters of the speech processing system according to the intensity of the noise quantity in order to reproduce a good quality speech. This paper presents a three layered neural network which is able to train the three graded speeches containing three graded white noise. Experimental results demonstrate that the noise amount could be discriminated by a neural network, whose inputs are cepstral coefficients. We believe that the improvement of speech signals degraded by noise is fully accomplished by attaching this discriminating system to the signal processing system.

Keywords: Speech enhancement, Noise reduction, Discriminating noise quantity, Speech recognition, Neural network

1 Introduction

Noise reduction and speech enhancement have been found many applications in speech recognition, aircraft communication, hearing aid and so on. In order to reduce the noise intensity in a conversation under the noisy environment, the methods by spectral subtraction[1, 2], wiener filter[3], microphone array[4, 5], and neural network[5, 6] have been developed in the past. The spectral subtraction method is necessary for the signal processing system to process adaptively according to the noise intensity in order to enhance the performance. For instance, in JEA S.LIM[1] the parameter "a" is chosen to be an appropriate value according to signal-to-noise ratio(SNR), so as to improve the speech intelligibility. According to Yan Ming Cheng[2], it is reported that the distortion measure of Itakura Saito is reduced by processing method I in lower SNR and by processing method II in higher SNR. In our case, the best amplitude adjustment coefficients also exist depending on the input SNR [7]. This paper presents a new idea to discriminate the noise quantity by a neural network and to improve the speech quality effectively using this result.

In our experiment, the neural network is trained by the speech in three levels: (1) the speech in noise-free, (2) the speech in light noise and (3) the speech in heavy noise, and the training result of the discriminating rates by the neural network is provided.

2 Outline of the noise quantity discrimination system

The discriminating noise quantity included in the speech signal, used as neural network of the perceptron type of a three layered structure, is trained by back propagation.

2.1 Construction of the system

The original speech signal is assumed to be $s(t)$, and the speech signal in which the noise is mixed is given by $x_k(t) = s(t) + k \times n(t)$. Here, $n(t)$ is a Gaussian white noise generated by a computer program with the sampling frequency of 8 kHz, where k is a coefficient of noise intensity which takes the values of 0, 3, and 6 in this research.

[1] Student, Doctor Course of Department of Information and Computer Engineering
[2] Lecturer, Department of Information and Computer Engineering
[3] Associate Professor, Department of Information and Computer Engineering
[4] Professor, Department of Information and Computer Engineering
[5] Professor, Department of Electrical Engineering

The schematic diagram of the experimental system is shown in Fig. 1. Discrete-time signal $x_k(t)$ of the sampling frequency 8 kHz is divided into the frames of 256 samples by square window $W(t)$, and effective value R_f of each frame are obtained and only the frames, of which R_f are smaller than threshold $T_h = r_m/3$, are used for discrimination. Here, r_m is an effective value obtained beforehand for each entire short sentences in order to simplify the experiment, but in practical applications r_m should be obtained by moving average. Here 10 or 20 cepstrums for each frame are adopted to the input of the neural network.

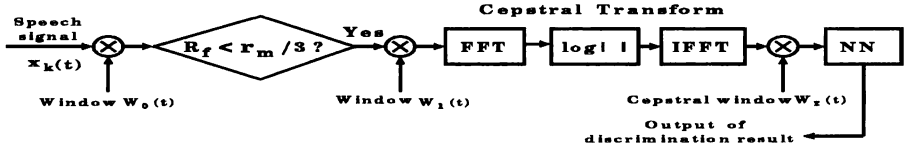

Fig. 1 Schematic diagram of the system

2.2 Original speech samples

The short sentences of which sampling frequency is 16 kHz, are made by the adult male speakers and the adult female speakers, and are used as the speech data, after the high region component has been removed through the low-pass filter of which cutoff frequency is 3.9 kHz, where the signals are decimated to 8 kHz sampling. The speech used are in total 12 kinds of the short sentences, which are composed of 6 kinds of the short sentences M1 $\sim$ M6 by the male speakers and 6 kinds of the short sentences F1 $\sim$ F4, F7 and F8 by the female speakers.

Table 1 shows the relationship between three male speakers and three female speakers and the short sentences. Training data for the neural network were the short sentences of M1, M2, F3, and F4, and the short sentences of the remain were used to discriminate the noise quantity. The bold letters show the training data used for the neural network.

Table 1 Relation between speech sentences and speakers

Male speakers	Short Sentences	Female speakers	Short sentences
Speaker 1	**M1, M2**	Speaker 1	F1, F2
Speaker 2	M3, M4	Speaker 2	**F3, F4**
Speaker 3	M5, M6	Speaker 3	F7, F8

3 Experimental setup and discrimination method

Spectrum envelope of the speech can be expressed completely by the part of low cepstrum. Only the sample of the frames, whose effective value R_f of each frame are smaller than threshold $T_h = r_m/3$ for the frame of 256 samples were used for the neural network training.

The neural network was composed of three layers, and the inputs of the neural network were 20 cepstrums for each frame and the composition of the neural network was 20-20-3. Target signals of the neural network were assumed that the state of noise-free was [1.0, -1.0, -1.0], the state of light noise was [-1.0, 1.0, -1.0], the state of heavy noise was [-1.0, -1.0, 1.0]. In this training, the training coefficient was assumed to be 0.1, the coefficient of inertia was assumed to be 0.5, and the training iteration was discontinued by 10,000 times. However, it was judged that the neural network had been settled by training of 10,000 times, because there were few error changes even if 10,000 times were exceeded.

The measure of performance evaluation of the system was represented as the amount of discrimination rate. The definition of this discrimination rate, which is a ratio of the number of frames correctly discriminated to the number of all frames used as training input, is shown in eq. (1).

$$\text{Discriminating rate}(\%) = \frac{\text{Number of frame in which the noise quantity was correctly discriminated}}{\text{Number of all frames given to input}} \times 100 \quad (1)$$

4 Results of the discrimination of noise quantity

The neural network was trained by using three kinds of speech data, which consisted of k = 0, 3, 6. The discrimination experiments of noise quantity were made by changing three parameters: (1) the threshold, (2) the short sentences, and (3) the speakers, for each of the speech (i) used to training and (ii) not used to training. The discrimination experiments of the noise quantity were also made for threshold $T_h = r_m/5$ and r_m. Table 2 shows average values of segmental SNR for the short sentences of each frame for reference[8]. The bold letters also show training data used for the neural network.

Table 2 Relation between k and SNR_{seg}

Short sentences	k = 3	k = 6	Short sentences	k=3	k=6
M1	**4.97dB**	**-1.05dB**	F1	9.16dB	3.14dB
M2	**5.04dB**	**-0.98dB**	F2	9.95dB	3.93dB
M3	4.13dB	-1.89dB	**F3**	**5.11dB**	**-0.91dB**
M4	3.83dB	-2.19dB	**F4**	**5.85dB**	**-0.17dB**
M5	9.91dB	3.89dB	F7	8.51dB	2.49dB
M6	5.78dB	-0.24dB	F8	9.01dB	2.99dB

4.1 Effect for noise discrimination performance of threshods

After the short sentences M1, M2, F3, and F4 were trained by the neural network, then the discrimination experiments were performed. Table 3 shows the discrimination rates when the thresholds were changed. The figures in parentheses show the rates of the number of frames of eq. (1). As shown in the table, it is found that the number of frames used for discrimination increases with the increment of the threshold. The discrimination rates of the noise quantity show high discrimination rate of 88% or more for each threshold. Fig. 2 shows the discrimination rates for three thresholds and their average values shown at the right edge. The discrimination rates were the highest for the threshold of $T_h = r_m/3$.

Table 3 Discrimination rates of the neural network based on the difference of T_h's(%)

A : In the case of $T_h = r_m/5$

Discrimination sentences	k = 0	k = 3	k = 6
M1	99.1(109/110)	99.1(109/110)	100(110/110)
M2	98.7(74/75)	100(75/75)	100(75/75)
F3	100(94/94)	97.9(92/94)	100(94/94)
F4	100(80/80)	98.8(79/80)	100(80/80)

B : In the case of $T_h = r_m/3$

Discrimination sentences	k = 0	k = 3	k = 6
M1	99.2(126/127)	100(127/127)	100(127/127)
M2	98.8(80/81)	100(81/81)	100(81/81)
F3	100(102/103)	100(103/103)	100(103/103)
F4	100(90/90)	100(90/90)	100(90/90)

C : In the case of $T_h = r_m$

Discrimination sentences	k = 0	k = 3	k = 6
M1	97.2(173/178)	96.1(171/178)	100(178/178)
M2	95.9(116/121)	91.7(111/121)	100(121/121)
F3	96.4(135/140)	90.7(127/140)	100(140/140)
F4	93.9(123/131)	88.6(116/131)	100(131/131)

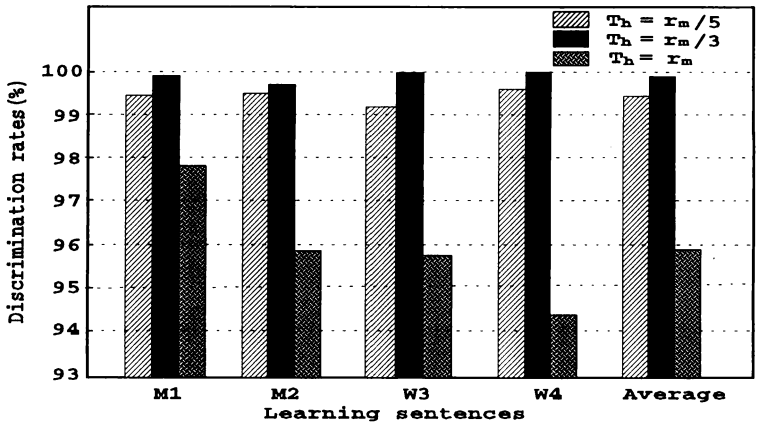

Fig. 2 Comparison of noise discrimination rates based on the difference of thresholds

4.2 Effect of speakers and short sentences on noise discrimination performance

The discrimination rates in the case of same sentences and different speakers are shown in Table 4. Table 5 shows the discrimination rates of different sentences and different speakers to the sentences of M5, M6, W7, and W8 after the neural network is trained by four kinds of the short sentences M1, M2, W3, and W4. In this case, the threshold value of $T_h = r_m/3$ was the most effective than the other thresholds. The average value in Table 3.B shows the discrimination rate of 99.9%, the average value in Table 4 shows the discrimination rate of 98.1%, and the average value in Table 5 shows the discrimination rate of 95.7%. The average values in any cases show a high discrimination rate of 95% or more, and it can be said that the effectiveness of the noise discrimination was proved to be possible by the neural network.

Table 4 Discrimination rates of the neural network based on the difference of speakers(%)
In the case of $T_h = r_m/3$

Training sentences	Discrimination sentences	k = 0	k = 3	k = 6
M1	F1	96.9	97.9	100
M2	F2	95.5	97.0	100
F3	M3	97.4	96.2	100
F4	M4	97.4	98.7	100

Table 5 Discrimination rates of the neural network based on the difference of speakers and sentences(%)
In the case of $T_h = r_m/3$

Training sentences	Discrimination sentences	k = 0	k = 3	k = 6
M1	M5	89.1	92.7	100
M2	M6	87.8	95.7	100
F3	F7	96.6	94.8	100
F4	F8	97.8	93.3	100

5 Application and examination

Table 6 shows the discrimination rates for the noise quantity of three levels k = 0, k = 3, and k = 6, when the noise of k = 0 ~ 7 is added to M6 and W7. As shown in the table, it is understood that k = 0, 1 are discriminated to k = 0, and k = 2, 3, 4 are discriminated to k = 3, and k = 5, 6, 7 to k = 6 at the highest rate. That is, the noise intensity added is appropriately discriminated to the suitable levels. Using this result we are intended to construct a speech enhancement system in which the parameters are adjusted according to the noise amount.

Table 6 Discrimination rates of the neural network for eight level noise quantities(%)
In the case of $T_h = r_m/3$

Training sentences	Noise quantity	Sentence M6			Sentence W7		
		k = 0	k = 3	k = 6	k = 0	k = 3	k = 6
M1	k = 0	**89.1**	10.9	0.0	**97.8**	2.2	0.0
	k = 1	**90.9**	9.1	0.0	**93.3**	6.7	0.0
M2	k = 2	45.5	**54.5**	0.0	37.8	**62.2**	0.0
	k = 3	1.8	**92.7**	5.5	2.2	**93.3**	4.5
F3	k = 4	0.0	**78.2**	21.8	0.0	**90.0**	10.0
	k = 5	0.0	12.7	**87.3**	0.0	11.1	**88.9**
F4	k = 6	0.0	0.0	**100**	0.0	0.0	**100**
	k = 7	0.0	0.0	**100**	0.0	0.0	**100**

6 Summary

Results achieved by the experiment above are as follows:

1. The discrimination of the noise quantity for the speech up to about 0 dB can be well performed by the neural network.
2. Different discrimination results are obtained for the different thresholds.
3. Good discrimination rates is obtained also for $T_h = r_m$ or $r_m/5$, though the case of threshold $T_h = r_m/3$ is the best.
4. The discrimination performance of 95% or more for the average is obtained though the speakers and the short sentences are different from the training data.

The research problems in future are (1) to confirm the effectiveness of this system for colored noise and (2) to apply the discrimination results to the speech enhancement system.

References

[1] JAE S. LIM: "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise", IEEE Trans. Acoust. Speech Signal Process., Vol.ASSP-26, No.5, pp.471–472, 1978.

[2] Yan Ming Cheng and Douglas O'Shaughnessy: "Speech Enhancement Based Conceptually on Auditory Evidence", IEEE Trans. Signal Processing, Vol.39, No.9, pp.1943–1953, 1991.

[3] Sreenivas and Pradeep Kirnapure: "Codebook Constrained Wiener Filtering for Speech Enhancement", IEEE Trans. Speech and Audio Processing, Vol.4, No.5, pp.383–389, 1996.

[4] Stephan Oh,Vishu Viswanathan and Panos Papamichalis: "Hands-Free Voice Communication in an Automobile with a Microphone Array", 1992 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92, pp.I-285–I-288, 1991.

[5] Wolfgang G. Knecht, Marks E. Schenkel, George S. Moschytz: "Neural Network Filters for speech Enhancement", IEEE Trans. Speech and Audio Processing, Vol.3, No.6, pp.433–438, 1995.

[6] S. Tamura: "An analysis of a noise reduction neural network", Proc. ICASSP Glasgow, Scotland, pp2001–2004, May, 1989.

[7] Jae-Seung Choi, Jiro Okamoto, Shoichi Hosokawa: "Improvement of the Characteristic of Speech which is buried in Noise based on Auditory Physiology", T.IEE Japan, Vol.115-C,No.11, pp.1332–1337, Nov., 1995.

[8] Kenzo ITOH, Nobuhiko KITAWAKI and Kazuhiko KAKEHI: "A Study of Objective Quality Measures for Digital Speech Waveform Coding Systems", IEICE Vol.J 66-A No.3, pp.274–281, 1983.