An Adaptive Speech Enhancement Method According to Noise Intensity Based on an Auditory Model

Jae-seung CHOI 1, Makoto HIRAI 2, Shigeyoshi NAKAJIMA 3, Shoichi HOSOKAWA 4, and Jiro OKAMOTO 5

Synopsis

In speech recognition under noisy environments, it is necessary to construct a system which reduces noise and enhances speech. It is effective to imitate the human auditory system which has an excellent mechanism to analyze the spectrum for speech enhancement. This paper presents an adaptive optimizing method applying the auditory mechanism which is called lateral inhibition. This method first estimates noise intensity by neural network, then adjusts adaptively the coefficients of both lateral inhibition and amplitude component according to the noise intensity for each input frame. It is confirmed that this method is effective for speech degraded not only by white noise but also by colored noise, judging from spectral distortion measurement.

Keywords: Neural network, Estimation of noise intensity, Lateral inhibition, Noise reduction, Adaptive speech enhancement

1 Introduction

Noise reduction is necessary not only for preprocessing of the speech recognition system but also for increment of intelligibility and reduction of auditory fatigue. This paper presents a speech enhancement system which reduces such troublesome noise.

Speech enhancement and noise reduction have been found to be useful in many applications in speech recognition, aircraft communication, hearing aids and so on. There are some approaches to reduce noise in conversation under noisy environments, such as spectral subtraction[1, 2, 3, 4, 5], Wiener filter[6], adaptive filtering[7, 8, 9], microphone array[10, 11], and neural network[11, 12, 13]. Spectral subtraction method is necessary for the signal processing system to process adaptively the speech according to the noise intensity in order to enhance performance. For instance, in J. S. Lim[1, 2] the parameter "a" is chosen to be an appropriate value according to signal-to-noise ratio (SNR), so as to improve the speech intelligibility. Moreover, the speech intelligibility is improved by choosing the length of the filter for the pitch period according to SNR. According to Y. M. Cheng et al., it is reported that the distortion measure of Itakura-Saito is reduced by processing method I in lower SNR, on the other hand by processing method II in higher SNR[3]. In our case, both the optimal adjustable coefficient \( B_f \) of lateral inhibition and \( R \) of amplitude component also exist depending on the amount of the input segmental SNR (\( SNR_{seg} \))[4].

On the other hand, according to the development of auditory physiology in recent years, numerous researches which imitate the mechanisms and functions of signal processing in auditory system are reported [3, 14, 15, 16, 17, 18]. We used one of these mechanisms called the lateral inhibition as a noise reduction system.

1 Student, Doctor Course of Department of Information and Communication Engineering
2 Associate Professor, Department of Information and Communication Engineering
3 Lecturer, Department of Information and Computer Engineering
4 Honorary Professor, Osaka City University
5 Professor, Department of Information and Communication Engineering
In general, the noise intensity is obtained by the signal intensity in the time domain which does not contain speech signal, but the detection of non-speech section from the speech signal containing noise is not easy\cite{19, 20}. As one of the methods to find a solution for the above-mentioned problem with simple composition, we propose an estimation method by a neural network (NN) under various noisy environments \cite{21}.

This paper first confirms the basic performance of the estimation method for the noise intensity we proposed, and then presents the performance of an adaptive speech enhancement system, which functions as an adaptive filter according to the noise intensity, that is, as the model of lateral inhibition mechanism in basal membrane of the inner ear. The whole speech enhancement system adjusts both $B_f$ and $R$ according to the noise intensity for each frame. As an evaluation criterion of improvement of speech enhancement, we used spectral distortion measurement (SD) which relates well to the speech intelligibility. It has been clarified that this method is effective for the speech enhancement not only for white noise but also for colored noise, judging from SD.

## 2 Experimental conditions

### 2.1 Speech signal

The original speech signal is assumed to be $s(t)$, and the speech signal disturbed by noise is given by $x_k(t) = s(t) + k \times n(t)$. Here, $n(t)$ is white noise or colored one with a sampling frequency of 8 kHz, and $k$ takes the values of 0, 3, and 6. Where colored noise is generated by passing white noise through low-pass filter with an amplitude-frequency characteristic as shown in Fig. 1. Since the spectrum of this colored noise looks extremely like the spectrum of real speech, it is thought that the estimation of noise from speech is not easy for NN so that noise reduction is difficult. However, it is shown in section 3 that the estimation of the noise intensity from the speech is not so difficult by the NN we proposed. The effective value of both white noise and colored one for $k=1$ is set to 0.33, where the average effective value for clean speech is 1.0. SNR$_{seg}$ is almost $-5$ dB for $k=6$ when the effective value of the noisy speech input is set to 1.0.

![Fig. 1 Amplitude-frequency characteristic of low-pass filter.](image)

### 2.2 Speech data

The original speech data, of which sampling frequency is 16 kHz, are of adult male speakers and of adult female speakers, their native language is Japanese. These are used as speech data, after the high region components are removed through low-pass filter of which cutoff frequency is 3.9
kHz, then the signals are decimated to 8 kHz sampling. The speech data used are a total of 6 kinds, composed of 3 kinds of the speech data M1 ~ M3 by the male speakers and 3 kinds of speech data F1 ~ F3 by the female speakers.

Table 1 shows the relationship between the speech data and speakers. The bold speech data of M1, M2, F1, and F2 are used to train the NN, and the remaining data are used to estimate the performance of the noise intensity detection. Table 2 shows the average values of $\text{SNR}_{seg}$ for each frame.

Table 1 Relationship between speech data and speakers.

<table>
<thead>
<tr>
<th>Male speakers</th>
<th>Speech data</th>
<th>Female speakers</th>
<th>Speech data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker 1</td>
<td>M1, M2</td>
<td>Speaker 1</td>
<td>F1, F2</td>
</tr>
<tr>
<td>Speaker 2</td>
<td>M3</td>
<td>Speaker 2</td>
<td>F3</td>
</tr>
</tbody>
</table>

Table 2 Relationship between $k$ and $\text{SNR}_{seg}$.

<table>
<thead>
<tr>
<th>Speech data</th>
<th>White noise addition</th>
<th>Colored noise addition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 3$</td>
<td>$k = 6$</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>$k = 6$</td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>0.79dB</td>
<td>-5.23dB</td>
</tr>
<tr>
<td>M2</td>
<td>0.76dB</td>
<td>-5.26dB</td>
</tr>
<tr>
<td>M3</td>
<td>0.36dB</td>
<td>-5.66dB</td>
</tr>
<tr>
<td>F1</td>
<td>0.60dB</td>
<td>-5.42dB</td>
</tr>
<tr>
<td>F2</td>
<td>0.59dB</td>
<td>-5.43dB</td>
</tr>
<tr>
<td>F3</td>
<td>0.22dB</td>
<td>-5.80dB</td>
</tr>
</tbody>
</table>

3 Estimation of the noise intensity by the NN

3.1 Neural network for estimating the noise intensity

In order to estimate the noise intensity embedded in noisy speech signal, the NN of perceptron type is used and is trained by back propagation as shown in Fig. 2.
Input and output for each unit approximates a sigmoid function, whose range of the output is from $-1$ to $+1$, as given by eq. (1).

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

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 input unit $i$ and output unit $j$, and $\theta$ is a threshold in each unit.

The NN is composed of three layers, and the inputs of the NN are 10 cepstrums for each frame in speech. The NN is trained by using three kinds of speeches: (1) noise-free speech ($k=0$), (2) speech with light noise ($k=3$) and (3) speech with heavy noise ($k=6$). Where, $k$ is a coefficient indicating the noise intensity. Table 3 shows the parameters used for implementation of the training and other conditions. The target signals for the NN is also shown in Table 4.

<table>
<thead>
<tr>
<th>Table 3 Various conditions for training of the NN.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial weight</strong></td>
</tr>
<tr>
<td>Coefficient of training</td>
</tr>
<tr>
<td>Coefficient of inertia</td>
</tr>
<tr>
<td>Construction of network</td>
</tr>
<tr>
<td>Maximum training iteration</td>
</tr>
<tr>
<td>Effective value of input</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4 Target signals for the NN.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise-free speech ($k=0$)</td>
</tr>
<tr>
<td>Speech with light noise ($k=3$)</td>
</tr>
<tr>
<td>Speech with heavy noise ($k=6$)</td>
</tr>
</tbody>
</table>

### 3.2 Construction of the noise intensity estimation system

The schematic diagram of the experimental system is shown in Fig. 3. The speech signal is first introduced the noise of intensity $k$ and synthesized as the discrete-time signal $x_k(t)$. Then it is transformed into cepstrum after passing a Hamming window $W_1(t)$ of 128 samples per frame. The obtained cepstrum is passed through window $W_2(t)$ and ten cepstrum components, consisting of from the 1st to the 10th of low region, are used as the input to the NN. Through these experiments, $x_k(t)$ is normalized to be a constant effective value for each clause of sentences. In normalization, the effective value of the speech added by the noise of $k=6$ is adopted as the normal standard, that is, one, and the effective values of all clauses of sentences are adjusted to this level.

![Fig. 3 Schematic diagram of the estimation system by the NN (NNES).](image-url)
3.3 Experiments of the noise intensity estimation

Estimation of the noise intensity \(k\) is made by the NN with the connection coefficients obtained after training. We evaluated the performance of this system by the amount of correct estimation rate. The definition of this estimation rate is the ratio of the number of frames in which the noise intensity is correctly estimated to the number of all frames given to the input, and it is shown in eq. (2).

\[
\text{Estimation rate(\%)} = \frac{\text{Number of frames in which the noise intensity is correctly estimated}}{\text{Number of all frames given to the input}} \times 100
\]

The estimation experiments of the noise intensity are made by changing two parameters: (1) the speech data, and (2) the speakers. For each data there are two kinds of the speech, that is, (i) used to training and (ii) not used to training. After the speech data M1, M2, F1, and F2 are trained by the NN, the estimation experiments are performed.

The estimation rates when the training speech data and the estimating ones are the same are shown in Table 5. In this Table, the estimation rates of the noise intensity show the high estimation rates of 95% or more for white noise and 92% or more for colored noise. Table 6 shows the estimation rates for M3 and F3 those are different from the training data, after the NN is trained by four kinds of the speech data M1, M2, F1, and F2. In this Table, the estimation rates of the noise intensity show the high estimation rates of 90% or more for both white noise and colored noise.

From the results shown in these Tables, it can be concluded that the noise estimation is effectively performed by the NN.

Table 5 Correct estimation rates when the training speech data and the estimating ones are the same.

<table>
<thead>
<tr>
<th>Estimating speech data</th>
<th>White noise addition</th>
<th>Colored noise addition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(k = 0)</td>
<td>(k = 3)</td>
</tr>
<tr>
<td>M1</td>
<td>100%</td>
<td>98.0%</td>
</tr>
<tr>
<td>M2</td>
<td>100%</td>
<td>95.2%</td>
</tr>
<tr>
<td>F1</td>
<td>100%</td>
<td>96.1%</td>
</tr>
<tr>
<td>F2</td>
<td>99.0%</td>
<td>95.0%</td>
</tr>
</tbody>
</table>

Table 6 Correct estimation rates when the speakers and the speech data are different.

<table>
<thead>
<tr>
<th>Training speech data</th>
<th>Estimating speech data</th>
<th>White noise addition</th>
<th>Colored noise addition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(k = 0)</td>
<td>(k = 3)</td>
<td>(k = 6)</td>
</tr>
<tr>
<td>M1, M2</td>
<td>M3</td>
<td>99.1%</td>
<td>90.4%</td>
</tr>
<tr>
<td>F1, F2</td>
<td>F3</td>
<td>99.4%</td>
<td>90.8%</td>
</tr>
</tbody>
</table>

4 Adaptive speech enhancement system

4.1 Spectral average

The intelligibility of the speech signal is very sensitive to components of spectral peaks. Therefore, unexpected peaks in the enhancement of heavy noisy speech which are generated by the sidelobes of the window function in a short-time power spectrum should not be generated, because they make musical noise[3, 22].
As a way to reduce unexpected peaks that are irregular from frame to frame, we adopt a spectral average as,

\[ \tilde{P}_k(i, \omega) = \frac{1}{2M + 1} \sum_{j=-M}^{M} W_j P_k(i - j, \omega). \]  

In this paper, we set \( M = 2 \) and \( W_{-2} = W_2 = 0.7, \ W_{-1} = W_1 = 1.1, \ W_0 = 1.4 \) concerning weights. Where \( \tilde{P}_k(i, \omega) \) is the short-time power spectral average of the \( i \)th frame and \( P_k(i, \omega) \) is an original spectrum before averaging.

### 4.2 Lateral inhibition

The function of spectral lateral inhibition (FSLI) is the imitation of the lateral inhibition of the nervous system in basal membrane of the inner ear, and it has the effect to emphasize the spectral peaks of speech and simultaneously compress the noise spectral valleys. Therefore, FSLI is an effective means for the speech enhancement [3, 4, 16, 17].

Fig. 4 shows impulse responses of three kinds of the FSLI, those should be selected according to the noise intensity. The horizontal axis shows frequency sampling points, and the vertical axis shows impulse response in the case when a unit input is added at a point \( B = 0 \), where \( B_f \) is a parameter which decides the width of the FSLI.

Parameters \( P_j \)'s show amplitude of impulse response and are restricted to satisfy eq. (4) for noise cancellation.

\[ P_l + P_c + P_r = 0 \]  

In this experiment, we set \( P_c = 1 \) and \( P_l = P_r = -0.5 \). Output of the FSLI \( B_k(i, \omega) \) is obtained by the convolution of \( \tilde{P}_k(i, \omega) \) with an impulse response shown in Fig. 4. In lateral inhibition, since the average value of the sum of the weighted noise is zero by this restriction, so the noise is reduced.
4.3 Construction of adaptive speech enhancement system

The construction of the proposed adaptive speech enhancement system is shown in Fig. 5. First, the speech signal degraded by the noise, sampled at 8 kHz, is normalized to a constant effective value per each sentence, and is transformed into cepstrum for each frame of 128 samples (above route). Then the noisy speech is delayed to 3 frames and is averaged by weighted spectral sum for each frame. The noisy signal $x_k(t)$ is added to the input whose level is the same as that of the training speech in order to function NNES in Fig. 3 correctly for input which contains unknown noise of different intensity. Spectral components obtained by the spectral average are convoluted with the FSLI in the frequency domain. The resulting amplitude spectrum after convolution may have some negative values. However, they do not have useful information in the present situation. So these are set to zero by rectifier (above route). $B_k(i, \omega)$ in the figure is the rectified result. On the other hand, the Fourier-transformed signal in another route is delayed to 3 frames, and is separated by amplitude component $A_k(i, \omega)$ and phase component $\theta_k(i, \omega)$. Finally, the enhanced speech signal $\hat{x}_k(t)$ is regenerated by IFFT using the following equation

$$F_k(i, \omega) = A_k(i, \omega)(1 + R \times B_k(i, \omega))e^{i\theta_k(i, \omega)}.$$  

(5)

Where $R$ is the amplitude coefficient, $i$ and $\omega$ represent frame number and spectral number, respectively.

5 Experiment of adaptive speech enhancement system

5.1 Estimation rates of the noise intensity for the NN

Table 7 shows the estimation results of the NN when the noise is adapted to the system in Fig. 3. Where, $\text{SNR}_{\text{seg}}$ varies from $\infty \text{dB}$ to $-7 \text{ dB}$, where $\infty$ means no noise is addition to speech. As shown in the table, it is understood that $k=0, 1$ is estimated to $k=0$, and $k=2, 3, 4$ is to $k=3$, and $k=5, 6, 7$ is to $k=6$ at the highest rate. That is, the noise intensity added is appropriately estimated to the suitable given levels. When this estimation system is adapted in the adaptive

![Figure 5: Adaptive speech enhancement system.](image-url)
speech enhancement system as shown in Fig. 5, for example, in the case when white noise added is 
k=4, then 4.1% of the frame is estimated to k=0, 65.3% to k=3, 30.7% to k=6, because SNRseg's 
differ in each frame. In this system, the parameters are adjusted to the optimal B_f and R for each 
frame according to these estimated results so as to enhance the speech.

Table 7 Estimation rates of the NN for k = 0 ~ 7 in the case when white noise and colored noise 
are added to speech data F3.

<table>
<thead>
<tr>
<th>Training speech</th>
<th>White noise SNRseg (dB)</th>
<th>Estimation rates</th>
<th>Colored noise SNRseg (dB)</th>
<th>Estimation rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k = 0</td>
<td>k = 3</td>
<td>k = 6</td>
<td>k = 0</td>
</tr>
<tr>
<td>M1</td>
<td>99.4</td>
<td>0.6</td>
<td>0.0</td>
<td>98.8</td>
</tr>
<tr>
<td>M2</td>
<td>75.1</td>
<td>17.4</td>
<td>7.5</td>
<td>73.3</td>
</tr>
<tr>
<td>F1</td>
<td>59.5</td>
<td>9.2</td>
<td>0.0</td>
<td>36.4</td>
</tr>
<tr>
<td>F2</td>
<td>90.8</td>
<td>5.2</td>
<td>0.0</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>65.3</td>
<td>30.7</td>
<td>0.0</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>76.3</td>
<td>20.2</td>
<td>0.0</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>92.4</td>
<td>5.9</td>
<td>0.0</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>94.2</td>
<td>1.7</td>
<td>0.0</td>
<td>1.8</td>
</tr>
</tbody>
</table>

5.2 Effect of adaptive speech enhancement based on SD

In the adaptive speech enhancement system shown in Fig. 5, parameters are adjusted to B_f=4 
and R=1.0 when the estimated result of the noise intensity is k=0, B_f=5 and R=2.0 when k=3, and 
B_f=6 and R=3.0 when k=6, for each frame. Fig. 6 shows spectral distortion (SD) of output \( \hat{x}(t) \) 
for white noise and colored noise. "R=0.0" in this figure shows SD for original speech degraded 
by white noise and colored noise, and "optimal R" shows the result obtained by adjusting only R 
to optimal value though B_f is fixed to 5, and "adaptive B_f, R" is SD obtained by our proposed 
method. According to the evaluated values of SD for white noise as shown in Fig. 6, SD is improved 
about maximal 6.3 dB and 9.1 dB, respectively, in order of "optimal R" and "adaptive B_f, R" 
to be measured from the basal level of "R=0.0". Moreover, a similar tendency can be found for 
colored noise in Fig. 6, that is, SD is improved about maximal 5.2 dB and 7.3 dB respectively. 
Moreover, as understood from Figure, the improvement of SD grows larger according to the noise 
intensity. The improvement of SD in "adaptive B_f, R" is about 3 dB for white noise and 2.6 dB 
for colored noise from "optimal R", respectively.

According to the experimental results, it can be said that the adaptive speech enhancement 
system by means of the estimation of the noise intensity is effective for both white noise and 
colored one.

Fig. 6 Effect of speech enhancement measured by SD when white noise and colored noise are added.
6 Summary

We proposed an adaptive speech enhancement system, which uses the model of lateral inhibition mechanism as a noise reduction filter adjusting parameters according to the noise intensity. In which we proposed the noise estimation system using the NN. Experimental results demonstrate that this system is effective for both white noise and colored one justifying from SD improvement. In summary, the results obtained by this experiment are as follows.

1. It is possible to make good estimation rates of the noise intensity by the NN for speech data up to about $-7 \text{ dB in } \text{SNR}_{\text{seg}}$.

2. Even if the speakers and the speech data are different from training data, the estimation rates of the noise intensity for both white noise and colored one by the NN are possible with accuracy of 90% or more on the average.

3. It is possible to improve the SD of noisy speech using optimal $B_f$ and $R$ which are the optimal estimated values per frame, in the adaptive speech enhancement system.

4. Effect of the noise reduction is large not only for colored noise but also especially remarkable for white noise.

The following problems are remained as the future problems.

1. Though we limited the estimating range of noise intensity to 3 levels as $k=0,3$ and 6, but we should examine whether it is necessary to divide the range into more than 3 levels.

2. To study whether this system is effective against such colored noise as road noise of real environment.

References


