Recognition of Stop Consonants Using Dynamic Critical Band Spectra and Neural Network

Hirofumi Yogo, Tadashi Kitamura and Naoki Inagaki
(Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466, Japan)

Abstract - Sophisticated procedures and fast analyzing method have been needed for analyzing the stop consonants. Because the features of the stop consonants are involved in the short transitional portion between the consonant and the vowel segments. An adaptive signal processing algorithm only needs a simple procedure. Moreover, a fast converging adaptive algorithm is obtained by utilizing the ASA method which is improved by authors. The estimated parameters are converted into the critical band spectra which are useful for recognizing speech. And, the critical band spectra are converted into dynamic critical band spectra with time-varying characteristics. These spectra are inputted into a three-layered neural network for recognizing the stop consonants. In this paper, an improving method of the learning properties of the neural network is also described.

1. Introduction
An intrinsic feature of stop consonants exists within a particular short portion of the speech waveform. Sophisticated procedures have been used for analyzing the stop consonants. We use an inverse system based on an adaptive digital filter (ADF) for analyzing the stop consonants. As an adaptive algorithm, we use the accelerated stochastic approximation (ASA) method.[1][2] Our accelerated stochastic approximation method was obtained by improving the convergence speed of a conventional stochastic approximation method. The ASA method has a fast convergence property and a simple algorithm. Therefore, it is suitable for analyzing a short duration data like the stop consonants. Generating process of the stop consonants is modeled by an autoregressive and moving-average (ARMA) process.[3][4]

In this paper, we obtained the spectra by calculating the FFT transformation of the estimated parameters. These spectra were converted into critical band spectra which are useful for recognizing the speech. An intrinsic feature of the stop consonants exists within the transitional portion between the consonant and the vowel segments. Therefore, the critical band spectra were calculated 14 times for that portion by shifting 2 mS. The obtained spectra were named as the ASA based dynamic critical band spectra. Recognizing the stop consonants was performed by inputting the spectra into a three-layered neural network. Moreover, the stop consonants were directly analyzed by using the FFT transformation for comparison, and the obtained spectra were also converted into the dynamic critical band spectra. The obtained spectra were named as the FFT based dynamic critical band spectra and were also inputted into the neural network. In this paper, improving the learning procedures of the neural network was also described.

2. ASA method
Speech production process of the stop consonants is modeled by an autoregressive and moving-average (ARMA) process. An inverse system shown in Fig. 1 was used for estimating the parameters of the model. An adaptive digital filter (ADF) shown in Fig. 1 is constructed by the ARMA model. As an adaptive algorithm of the ADF, the accelerated stochastic approximation (ASA) method was utilized. The ASA method has the fast converging property which is obtained by using the optimized convergence factor for each AR and MA parameter, respectively. An example of the
converging properties is shown in Fig. 2. Less than 25 iterations is needed to converge. The speech parameters are estimated by using Eq. (1).

\[ G(j+1) = G(j) + e(j) \mu(j) P(j) / (j + h) \]  

(1)

where,

\[ G(j) = [b_0(j), ..., b_N(j), a_1(j), ..., a_M(j)] \]

\[ P(j) = [\beta_0(j), ..., b_N(j), a_1(j), ..., a_M(j)] \]

\[ e(j) = d(j) - W(j)^T \Sigma(j) \]  

where, \( \Sigma(j) \) is a diagonal matrix containing the respective variances of the input signal. \( d(j) \) is the signal at time \( j \), and \( W(j) \) is a weight vector.

\[ \mu(j) = \frac{\partial e(j)}{\partial a(j)} \]  

(2)

\[ \beta_p(j) = \frac{\partial e(j)}{\partial b_p(j)} \]  

(3)

\[ \alpha_k(j) = \frac{\partial e(j)}{\partial a_k(j)} \]  

(4)

\[ \beta_p(j) = \frac{\partial e(j)}{\partial b_p(j)} \]  

(5)

\[ \alpha_k(j) = \frac{\partial e(j)}{\partial a_k(j)} \]  

(6)

\[ \mu^{*}(j) = 0.5 \alpha \frac{\partial e(j)}{\partial a_k(j)} \]  

(7)

where,

\[ \alpha = (M + N) / (M + N + 1), \ \mu_0 = 1 / (M + 1) \sigma^2 \]

\[ ||\cdot||: \text{vector norm}, \ n, c, \gamma: \text{constant}, \ \sigma^2: \text{input power} \]

3. Parameter estimation and the critical band spectra of the stop consonants

As the speech data of the stop consonants, the JEIDA'S Japanese common speech data base was used, sampled by 12 kHz and converted into 12 bits. The parameter estimation and converting the estimated parameters into the critical band spectra were performed as shown in Fig. 3. The estimated parameters were FFT(256 points) transformed and converted into the critical band spectra by averaging.
Fig. 4 An example of the dynamic critical band spectra (Unvoiced stop consonant, from left to right: /p, t, k/) upper side: ASA based ones, lower side: FFT based ones.

Fig. 5 An example of the dynamic critical band spectra (Voiced stop consonant, from left to right: /b, d, g/) upper side: ASA based ones, lower side: FFT based ones.
them in each critical band shown in Table 1 [5]. This operation was repeated 14 times for the transitional portion between the consonant and the vowel segments, and ASA based dynamic critical band spectra with time-varying characteristics were obtained. The same operation was also performed by directly applying the FFT transformation to the speech waveform, and the FFT based dynamic critical band spectra were obtained. An example of the spectra of the unvoiced stop consonants obtained by these operations is shown in Fig. 4(a) and (b). Moreover, an example of the ones of the voiced stop consonants is shown in Fig. 5(a) and (b). In the both Figure 4 and 5, the ASA based dynamic critic band spectra shown in the upper side have more fine and stable features than the FFT based dynamic critical band spectra shown in the lower side.

4. Improving the Learning Properties of the Neural Network

A three-layered neural network shown in Fig. 6 was used for recognizing the stop consonants. The input layer has 198 units (=18 x 11) for accepting the spectra shown in Fig. 4 or Fig. 5, and the output layer has 3 units. The optimum unit number of the hidden layer was experimentally fixed, as described later on. The sigmoid function given by Eq. (8) is generally used as the nonlinear function.

\[
\sigma_i = \frac{1}{1 + \exp\left(-\sum w_{ji} o_j + \gamma_i + B\right)}
\]

where,
\[\gamma_i : \text{threshold of each unit}, \quad o_i : \text{output of the unit i}, \quad w_{ji} : \text{weight between the } j\text{th and the} \quad i\text{th units}, \quad B : \text{common bias of each unit}\]

After setting the unit number of the hidden layer to 20, learning procedures with the back propagation (BP) algorithm were performed by changing the bias B, and the learning properties shown in Fig. 7 were obtained. The optimum value of B is observed to be B=0.2 in the figure. In the next, the learning procedures were repeated by changing the momentum weight \(\alpha\) in Eq. (9) after setting B=0.2, and the learning properties shown in Fig. 8 were obtained.

\[
\Delta w_{ji}(n+1) = \eta \delta_{pj} o_{pj} + \alpha \Delta w_{ji}(n)
\]

where,
\[\eta : \text{learning rate}, \quad \alpha : \text{momentum weight}, \quad \delta_{pj} : \text{error between the output and the desired training value}, \quad o_{pj} : \text{output value}, \quad n : \text{learning time index}\]

When the \(\eta\) was 0.4, the characteristics with less errors and stable convergence property were obtained. The optimum values of the bias B, the learning rate \(\alpha\), and the momentum weight \(\eta\) are \(B=0.2, \quad \alpha=1.0, \quad \eta=0.4\), respectively. As the teaching data, four sets of (true, false) patterns were used, that is, (1) T1(0.9,0.1), (2) T2(0.8,0.2), (3) T3(0.75,0.25), and T4(0.7,0.3). The learning properties shown in Fig. 10 were obtained. The pattern T3 has the lowest learning errors. The optimum unit number of the hidden layer was determined as follows. After setting each parameter to the optimum value, learning procedures were repeated by changing the unit numbers from 18 to 30, and the recognition rates for the learning data were calculated. As the results, the recognition

<table>
<thead>
<tr>
<th>Critical Band (Bark)</th>
<th>Frequency (Hz)</th>
<th>Range (Hz)</th>
<th>Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20 ~ 200</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>200 ~ 300</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>300 ~ 400</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>400 ~ 510</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>510 ~ 630</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>630 ~ 770</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>770 ~ 930</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>920 ~ 1080</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1080 ~ 1270</td>
<td>190</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1270 ~ 1480</td>
<td>210</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1480 ~ 1720</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1720 ~ 2000</td>
<td>280</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2000 ~ 2320</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2320 ~ 2700</td>
<td>380</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2700 ~ 3150</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3150 ~ 3700</td>
<td>550</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>3700 ~ 4400</td>
<td>700</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>4400 ~ 5000</td>
<td>800</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Critical band and its bandwidth

It is found that the optimum value of \(\alpha\) is 1.0 from the results. Furthermore, after setting B=0.2 and \(\alpha=1.0\), the learning procedures were performed by changing the momentum weight \(\eta\) in Eq. (9), and the properties shown in Fig. 9 were obtained.
Fig. 6 Derived configuration of the three-layered neural network.

Fig. 7 Learning properties when bias B is changed.

Fig. 8 Learning properties when momentum weight \( \alpha \) is changed.

Fig. 9 Learning properties when learning rate \( \eta \) is changed.

Fig. 10 Learning errors vs. learning iterations for several teaching pattern sets.

Fig. 11 Optimum hidden unit number
rates shown in Fig.11 were obtained. Therefore, it is found that the optimum unit number of the hidden layer is 26 with a margin.

5. Recognition experiment

The learning operation was performed by changing the number of the data frame to be inputted into the network from 5 to 14, prior to the recognition experiments. The learning errors were obtained as shown in Fig.12. More than 7 frames are needed for obtaining the excellent learning results. Therefore, in the recognition experiment, 11 frames were used with a margin. The unit number of the hidden layer, the bias B, the momentum weight \( \eta \) and the learning rate \( \alpha \) were set to the optimum values, and the learning of the neural network was performed. Ten male speaker's speech data of the JEIDA's Japanese common speech data base were utilized for the learning. The recognition experiments were successfully carried out for the ASA based and the FFT based dynamic critical band spectra, respectively. The recognition rate of the unvoiced stop consonants for the ASA based dynamic critical band spectra was 98.3 \%, and the rate for the FFT based ones was 90.0 \%. The recognition rate of the voiced stop consonants for the ASA based dynamic critical band spectra was 98.3 \%, and the rate for the FFT based ones was 96.7 \%.

6. Discussions and Conclusions

The learning properties of the neural network were greatly improved by using the common bias B for all units. The ASA based dynamic critical band spectra have more fine characteristics than the FFT based ones. The higher recognition rates for the ASA based dynamic critical band spectra show the usefulness of our proposed method. The sophisticated algorithm has been needed for estimating the parameters of the ARMA model.[3][4] However, the inverse system based on the proposed ASA algorithm needs only simple algorithm. The applications of the proposed method to the other phonemes are intended hereafter.

References