Improved Forward Masking on a Generalized Logarithmic Scale for Robust Speech Recognition

Hiroshi Matsumoto, Takumasa Ichikawa and Kazumasa Yamamoto

Department of Electrical & Electronic Eng., Faculty of Engineering, Shinshu University
4-17-1 Wakasato, Nagano-shi, Nagano 380-8553, Japan,
matsu@sp.shinshu-u.ac.jp

Abstract

We previously proposed a forward masking on a generalized logarithmic scale to eliminate convolutional noise as well as to suppress additive noise. While the generalized Dynamic Cepstrum derived from the masked spectrum has been robust to both noises, the robustness to convolutional noise slightly degrades as compared to masking on the logarithmic scale, and the optimal masking coefficient depends on SNR and the type of noises. This paper improves these issues by applying the variance normalization and by controlling the masking the level depending on the estimated SNR. The recognition tests using the Aurora2 database shows that the variance normalization improves the word accuracy even for the test set C, which includes MIRS distortion, from 79.2% for the logarithmic scale to 84.0% for the generalized logarithmic scale, and that the two level masking depending on SNR improves the word accuracy for speech babble noise of 10dB from 83.7% to 89.9%.

1. Introduction

In real environment, the speech recognition performance is severely degraded by mismatch between the training and testing environments. These mismatch may result from additive background noise and/or convolutional noise due to the frequency characteristics of microphones, etc. A variety of compensation techniques have previously been proposed for the front end and the decoder in speech recognition systems.

As to the front end, the dynamic cepstrum based on the forward masking on logarithmic spectral domain is known to be robust to additive noise as well as convolutional noise. This paper proposes a forward masking on the generalized logarithmic scale between power spectrum and logarithmic spectrum, and examines the effectiveness for applying it to mel frequency BPF spectrum.

Dynamic-cepstrum(DyC)[1] has been shown to be a powerful technique for additive and convolutional noise. DyC which simulates the forward masking intends to enhance transitional spectral features such as formant transitions, while suppressing time-invariant spectral characteristics. However, since DyC is based on subtraction of a masking pattern from a current spectrum on a logarithmic scale in amplitude, the compensation of additive noise is less robust than the compensation of convolutional noise.

In order to improve the robustness of DyC to additive noise while maintaining the robustness to convolutional noise, we examined the forward masking on the generalized logarithmic scale[DyMFGC][5][4] in amplitude. Since the generalized logarithmic spectra for $\gamma = 0$ and $\gamma = 1$ correspond to the the logarithmic and the power spectra respectively, an intermediate value of $\gamma$ is expected to make a compromise between both effects to additive and convolutional noise. The proposed forward masking on the generalized logarithmic scale is incorporated into MFCC analysis together with a equal loudness weighting as in PLP analysis[7], and this spectrum simulates psychophysics of hearing to derive an estimation of the auditory spectrum.

Moreover, since an insertion error rate changes with the values of a masking coefficient depending on the kind of noise, we propose a two-level masking method in which the masking coefficient is changed between voiced section and silent section. Furthermore, variance normalization is performed as post-processing. These combination is compared with other similar conventional techniques such as, Dynamic-cepstrum(DyC), continuous spectral subtraction method(CSS)[6], cepstral mean subtraction(CMS)[3] through connected digit recognition tests using the Aurora2 database[9].

2. Forward masking on the generalized logarithmic scale

This section overviews the procedure of the forward masking on the generalized logarithmic scale including the improved step and the post processing by variance normalization.

2.1. Equal Loudness Weighting

The mel-scale power spectrum is multiplied by a fixed equal-loudness weighting

$$X(n, k) = Y(n, k) \cdot E(k)$$

where $Y(n, k)$ is the power of the k-th filterbank channel at the n-th frame, and $E(k)$ is equal loudness weighting. This property of equal loudness curve simulates the frequency dependent sensitivity of hearing at the 40dB SPL.

2.2. Generalized Logarithmic Scale Conversion

Generalized logarithmic scale in amplitude is defined by

$$s_\gamma(\omega) = \begin{cases} \frac{(\omega^\gamma - 1)}{\gamma}, & \gamma \neq 0 \\ \ln\omega, & \gamma = 0. \end{cases}$$

The mel-scaled power spectrum is compressed by the generalized logarithmic function in amplitude as

$$X_\gamma(n, k) = S_\gamma\{X(n, k)\}.$$
Equation (2) shows that generalized logarithmic spectra for \( \gamma = 0 \) and \( \gamma = 1 \) correspond to the logarithmic and the power spectra respectively. Furthermore, this \( \gamma \)-th power operation is considered as an approximation to the power law of hearing which represents the nonlinear relationship between the intensity of sound and its perceived loudness.

### 2.3. Forward Masking and its Modification

The masked spectrum is formulated as the spectrum suppressed by the masker \( M(\gamma, n, k) \) from the current spectrum \( X(\gamma, n, k) \)

\[
P_{\gamma,\alpha}(n, k) = X(\gamma, n, k) - \alpha M(\gamma, n, k) \tag{4}
\]

where \( n, k \), and \( \alpha \) denotes the frame number, the channel number, and a masking coefficient, respectively. The masked spectrum \( M(\gamma, n, k) \) is the exponentially weighted sum of preceding spectra defined by

\[
M(\gamma, n, k) = \sum_{m=1}^{\infty} X(\gamma, n-m, k)\beta^m / \sum_{m=1}^{\infty} \beta^m \tag{5}
\]

where \( \beta \) is a decay rate. \( M(\gamma, n, k) \) is recursively calculated using the following Equation.

\[
M(\gamma, n, k) = \beta M(\gamma, n-1, k) + (1 - \beta)X(\gamma, n-1, k) \tag{6}
\]

In the above formulation [5], the masking coefficient \( \beta \) is fixed to a global optimal value for all the environmental conditions. However, since the optimal value depends on SNRs for some noise environments, the masking coefficient is modified to be SNR-dependent one. This will be described in the later section.

### 2.4. DCT and Gain Normalization

The masked spectrum \( P_{\gamma,\alpha}(n, k) \) is converted to the generalized cepstral coefficient \( C_{\gamma,\alpha}(n, i) \) by DCT, which is referred to as the dynamic mel-frequency generalized cepstrum (DyMFGC). Since the generalized cepstrum is not gain invariant, DyMFGC is finally converted to the gain normalized DyMFGC, \( \tilde{C}_\gamma(n, i) \) as

\[
\tilde{C}_{\gamma,\alpha}(n, i) = \begin{cases} 
C_{\gamma,\alpha}(n, 0) \cdot \tilde{X}(n)\gamma & i \neq 0 \\
C_{\gamma,\alpha}(n, 0) \cdot \tilde{X}(n)\gamma - S_i(\tilde{X}(n))(1 - \alpha) & i = 0
\end{cases} \tag{7}
\]

where \( \tilde{X}(n) \) is the mean spectral level at time \( n \) defined by

\[
\tilde{X}(n) = \left( \frac{1}{N} \sum_{i=1}^{N} X(n, k) \right)^{1/\gamma}. \tag{8}
\]

The gain normalized DyMFGC, \( \tilde{C}_\gamma(n, i) \) equals to the DCT coefficient of the generalized logarithmic spectra \( s_{\gamma}(X(m, k)/X(n)) \).

### 2.5. Variance Normalization

On the final stage of the front-end, Variance Normalization [8] is applied to DyMFGC \( C_{\gamma,\alpha}(n, i) \) such that their variance equals to unity.

\[
\tilde{c}_\gamma(n, i) = \tilde{c}_\gamma(n, i)\sigma_i^{-1} \tag{9}
\]

The variance is calculated from a utterance by

\[
\sigma_i^2 = \frac{1}{N} \sum_{n=1}^{N} (\tilde{c}_\gamma(n, i) - \bar{c}_\gamma(i))^2. \tag{10}
\]

This normalization is done on a utterance by utterance basis for both training and test data.

### 3. Evaluation of parameter

#### 3.1. Speech data and experimental conditions

The forward masking on the generalized logarithmic scale is evaluated through connected digit recognition tests. The speech data is from Aurora2 database with additive and convolutional noises. Models are trained by only clean speech with the frequency transfer characteristic of G.717 in this paper. Test speech in Aurora2 database is classified into three sets, Set A, B, and C. In this study, however, test data was reorganized into Group 1, 2, and 3 so that the environmental conditions in each group show the similar tendency on recognition errors as a function of the parameters in DyMFGC. Group1 is composed of additive noises (Subway, Car, Street, and Train-Statin) and the same convolutional noise (G.717) as the training speech. Group2 is composed of different noises from Group1 (Babble, Exhibition, Restaurant, and Airport) and the same convolutional noise (G.717) as Group1. Group3, the same as the original SetC, is composed of additive noises (Subway, and Street) and the different convolutional noise (MIRS) from the training speech.

Analysis condition and the structure of HMMs are shown in Table 1. Recognition performance is evaluated by the word accuracy as

\[
Acc[\%] = \frac{N - (S + I + D)}{N} \times 100 \tag{11}
\]

where \( N \) represents tokens, and \( S, D, \) and \( I \) are substitution, deletion, and insertion errors respectively.

<table>
<thead>
<tr>
<th>Analysis window</th>
<th>Frame Shift</th>
<th>Mel Filter Bank Channel</th>
<th>Feature vector</th>
<th>20ms Hamming</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>State</td>
<td>Mixture</td>
<td>left-to-right model</td>
<td>silent model : 3 states word model : 16 states 3 mixture</td>
</tr>
</tbody>
</table>
3.2. Effect of VN on the Masking Parameters

3.2.1. Effect of VN on $\gamma$

First we examined the optimal values of $\gamma$ at the masking condition of $\alpha = 0.6$ and $\beta = 0.8$ with/without variance normalization. SNR of all the test speech was set to 10dB. Figure 1 shows the word accuracies for three groups as a function of $\gamma$ with/without VN. From this figure, the variance normalization improves recognition accuracies between $\gamma = 0$ and 0.2 for all the test groups. Without VN the generalized logarithmic scale degrades recognition accuracies for Group3, which has a mismatch on the frequency characteristics. With VN, however, the value of $\gamma = 0.1$ gives highest recognition accuracies for all of the groups. Thus, the value of $\gamma$ will be set to 0.1 throughout the following experiments.

3.2.2. Effect of Masking Parameters

The second experiment examines the optimal values of the masking coefficient $\alpha$ with the fixed value of $\gamma = 0.1$ and $\beta = 0.8$. Figure 2 shows the recognition accuracies as a function of $\alpha$. With variance normalization, the optimal values of $\alpha$ tend to shift to smaller values, especially for Group1, and range from 0.4 to 0.8 depending on environmental conditions. In the subsequent experiments, the value of $\alpha$ is fixed to 0.6. Then, figure 3 shows the effect of the decay rate $\beta$ with the fixed values of $\gamma = 0.1$ and $\alpha = 0.6$. The optimal values of $\beta$ are between 0.7 and 0.98 for three groups, but the recognition accuracies are insensitive to $\beta$ except for Group3. Thus, the value of $\beta = 0.9$ will be used throughout the following experiments.

4. Two-level Masking Method

First, we examined the three types of errors, i.e., substitution, deletion, and insertion errors. As a result, it was found that large insertion errors occur significantly in Babble, Restaurant and Airport environment. The effect of the masking coefficient on three types of errors was further investigated. Figure 4 shows the error rates of insertion (INS) and sum of substitution and deletion (S+D) as a function of $\alpha$ under Babble noise conditions of 5dB and 20dB SNR. From this figure, it is found that while the error rate of S+N is minimum at $\alpha = 0.6$, the insertion error monotonically decrease as $\alpha$ increases to 1.0. Further examination of insertion errors revealed that it happens more frequently at lower SNR parts of speech. Taking account of these phenomena, we propose a two-level masking in which the value of $\alpha$ is set to 0.6 for higher SNR and to 0.98 for lower SNR. This switching between two values are controlled by the ratio $S(n)$ of the noisy power spectrum $X(n)$ to estimated noise power spectrum $N(n)$ as

$$S(n) = \frac{X(n)}{N(n)}$$

(12)

$$\alpha(n) = \begin{cases} 0.98 & S(n) < S_{th} \\ 0.6 & S(n) > S_{th} \end{cases}$$

(13)

The noise power spectrum was estimated by averaging 5 frames before speech onset. In order to determin an appropriate value of $S_{th}$, recognition tests was carried out for various values of $S_{th}$. Figure 5 shows the recognition accuracies as a function of $S_{th}$ for Babble and Car noise conditions of 10dB SNR. Here, $S_{th} = 0.0$ and $S_{th} = \infty$ correspond to fixing $\alpha$ to 0.6 and 0.98, respectively. Although the optimal thresholds for both noise conditions are different, the values of $S_{th}$ between 1.0 and 2.5 improves the word accuracy as compared to the fixed value of 0.
Figure 4: Effect of $\alpha$ on insertion error and sum of substitution and deletion errors under Babble noise condition.

Figure 5: The effect of $Sth$ on word accuracy.

$\alpha = 0.6$, i.e., $Sth = 0$. For all over the test sets, it was found that a lower threshold tends to be optimal for lower SNR. Thus, the value of $Sth$ is set to 1.0 in the following experiments.

5. Comparison with conventional methods

The robustness of our proposed method $DyMFGC_{SN}$ with two level masking was compared with DyMFC, which is the original forward masking on logarithmic scale, MFCC, and MFCC+CSS+CMS, which is MFCC with the continuous spectral subtraction as well as cepstral mean subtraction. Moreover, every feature vector contains delta cepstrum and is processed by the variance normalization. Table 2 shows the average word accuracies over Aurora2 test sets A, B, and C. The proposed method ($DyMFGC_{SN}$) outperforms the previous DyMFGC and DyMFC over all the SNRs, and achieved the highest performance among four feature parameters examined except at 0dB SNR.

7. References


Table 2: Comparison with conventional methods.

<table>
<thead>
<tr>
<th>Metric</th>
<th>0dB</th>
<th>5dB</th>
<th>10dB</th>
<th>15dB</th>
<th>20dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>29.20</td>
<td>55.75</td>
<td>75.10</td>
<td>86.03</td>
<td>90.91</td>
</tr>
<tr>
<td>MFCC+CSS+CMS</td>
<td>49.71</td>
<td>73.08</td>
<td>86.29</td>
<td>92.30</td>
<td>95.26</td>
</tr>
<tr>
<td>DyMFC</td>
<td>40.05</td>
<td>67.56</td>
<td>83.39</td>
<td>91.17</td>
<td>94.76</td>
</tr>
<tr>
<td>DyMFGC</td>
<td>40.12</td>
<td>71.40</td>
<td>87.07</td>
<td>93.49</td>
<td>96.04</td>
</tr>
<tr>
<td>DyMFGC_{SN}</td>
<td>44.74</td>
<td>74.04</td>
<td>88.30</td>
<td>93.80</td>
<td>96.16</td>
</tr>
</tbody>
</table>

The variance normalization made the generalized logarithmic scale of $\gamma = 0.1$ effective under all of the environmental conditions. Furthermore, the two level masking depending on SNR reduced the insertion error especially for Babble and Car noise conditions. As a result, the proposed method ($DyMFGC_{SN}$) attained the highest recognition performance over the previous feature parameters compared except at 0dB SNR.

6. Conclusions

This paper has presented the improved forward masking on the generalized logarithmic scale. First, the postprocessing with the