FORWARD MASKING ON A GENERALIZED LOGARITHMIC SCALE FOR ROBUST SPEECH RECOGNITION

Yoshihiro Ito, Hiroshi Matsumoto, and Kazumasa Yamamoto
Dept. of Electrical & Electronic Eng., Faculty of Engineering, Shinshu University
4-17-1 Wakasato, Nagano-shi, Nagano 380-8553, Japan,
E-mail: matsu@sp.shinshu-u.ac.jp

ABSTRACT

This paper examines the forward masking on the generalized logarithmic scale for robust speech recognition to both additive and convolutional noise. The forward masking in the dynamic cepstral (DyC) representation is based upon subtraction of a masking pattern from a current spectrum on a logarithmic spectral domain, whereas the proposed method intends to make a compromise between the logarithmic and linear spectral domains by choosing an appropriate value of the power. This technique is incorporated into a modified MFCC-based frontend. The connected-digit recognition tests showed that in noisy conditions this technique outperforms the conventional techniques such as the DyC, the continuous spectral subtraction method, the cepstral mean subtraction while maintaining the robustness to the convolutional noise.

1. INTRODUCTION

The stochastic-based speech recognizers severely degrade the performance by the mismatch between the training and testing environments. These mismatch may result from the additive background noise and/or the convolutional noise such as microphone frequency characteristics. A variety of compensation techniques and robust spectral parameters have previously been proposed. Among these techniques, the cepstral mean subtraction (CMS) [1], the relative spectral (RASTA) technique [2], and the dynamic cepstrum (DyC) [3] have been shown to be effective not only for convolutional noise but also for additive noise in some extent. However, the effectiveness to additive noise are limited since these are based upon a sort of a filtering in the logarithmic spectral domain.

In order to improve the robustness of the DyC to additive noise while maintaining the robustness to convolutional noise, this paper proposes the forward masking on the generalized logarithmic spectral domain [4]. On the basis of the forward masking concept, the DyC intends to enhance transitional spectral features such as formant transitions while suppressing time-invariant spectral characteristics.

However, since the DyC is based on subtraction of a masking pattern from a current spectrum on a logarithmic scale in amplitude, it is not appropriate for suppressing additive noise. The generalized logarithmic function is able to realize an intermediate amplitude compression between the logarithmic and linear functions depending on the value of the power. Therefore, it is expected to make a compromise between both effects to additive and convolutional noise. A similar technique, the Lin-Log RASTA [2], was also proposed. This method based upon the linear approximation of the logarithmic function, unlike the generalized logarithmic function.

The proposed method is applied to the mel-frequency filterbank spectral domain together with a equal loudness weighting as in the PLP analysis[5]. The recognition performance of forward masking on the generalized logarithmic scale is compared with other similar conventional techniques such as the DyC, the CMS, and the continuous spectral subtraction method (CSS) [6] through gender-independent connected digit recognition tests.

2. FORWARD MASKING ON THE GENERALIZED LOGARITHMIC SCALE

The forward masking for the DyC is simulated in the linear predictive spectral domain on the linear frequency scale, whereas the proposed forward masking is applied to the mel-scale filterbank spectra. Then, the masking process is incorporated into the modified MFCC analysis configuration.

(1)Mel-Scale Filterbank Analysis

The power spectrum of a windowed speech segment is obtained by the same mel-scale triangular filters used in the conventional MFCC analysis. In the following steps, \( Y(n, k) \) denotes the power of the \( k \)th channel at time \( n \) \((k = 1, \ldots, N; \ n = 0, 1, \ldots)\).

(2)Equal Loudness Preemphasis

In this study, the power spectrum \( Y(n, k) \) is preemphasized
by the simulated equal loudness curve $E(k)$ as in the PLP analysis [5].

$$X(n, k) = E(k)\hat{Y}(n, k)$$

(1)

This property of $E(k)$ simulates the relative sensitivity of hearing at the center frequency of the 4th channel at the 40dB level. While this preemphasis has no effect on the performance of a MFCC-based ASR system, the preemphasis combined with the generalized logarithmic amplitude conversion is expected to enhance the performance.

(3) Generalized Logarithmic Scale Conversion

The magnitude of $X(n, k)$ is compressed by the generalized logarithmic function $s_\gamma(w)$ [4] as

$$X_\gamma(n, k) = s_\gamma(X(n, k))$$

(2)

where $s_\gamma(w)$ is defined by

$$s_\gamma(w) = \left\{ \begin{array}{ll} (w^\gamma - 1)/\gamma & 0 < |\gamma| \leq 1 \\ w, & \gamma = 0. \end{array} \right.$$ 

(3)

This $\gamma$th power operation in equation (3) intends to approximate the power law of hearing as in the PLP analysis. The generalized logarithmic power spectra $X_\gamma(n, k)$ for $\gamma = 0$ and $\gamma = 1$ correspond to the logarithmic and the power spectra, respectively.

(4) Forward Masking

In the DyC representation, the forward masking is formulated as the current spectrum compressed by the masker spectrum, which is the sum of the preceding frequency-smoothed and decayed spectra [3]. In this study, however, the frequency-smoothing was excluded since a preliminary experiment has found that the smoothing slightly degrades the recognition performance for the mel-scaled filterbank spectrum. The masked spectrum $P_{\gamma,\alpha}(n, k)$ is then represented by

$$P_{\gamma,\alpha}(n, k) = X_\gamma(n, k) - \alpha M_\gamma(n, k),$$

(4)

where $\alpha$ is a subtraction coefficient, and $M_\gamma(n, k)$ is the masker spectrum. $M_\gamma(n, k)$ is the exponentially weighted sum of preceding spectra, and is recursively computed by

$$M_\gamma(n, k) = \beta M_\gamma(n-1, k) + (1 - \beta)X_\gamma(n-1, k)$$

(5)

where $\beta$ is a decay rate.

(5) DCT and Gain Normalization

The masked spectrum $P_{\gamma,\alpha}(n, k)$ is converted to the generalized cepstral coefficient $c_\gamma(n, i)$ by DCT, which is referred to as the dynamic mel-frequency generalized cepstrum (DyMFCC). Unlike the standard cepstrum, the generalized cepstrum is not free from the gain factor of speech. Therefore, DyMFCC is finally converted to the gain normalized DyMFCC, $\tilde{c}_\gamma(n, i)$ as

$$\tilde{c}_\gamma(n, i) = \left\{ \begin{array}{ll} c_\gamma(n, i)\hat{X}(n)^{\gamma} - s_\gamma(\hat{X}(n)) (1 - \alpha), & \gamma = 0 \\ c_\gamma(n, i)\hat{X}(n)^{\gamma}, & \gamma \neq 0 \end{array} \right.$$ 

(6)

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

$$\hat{X}(n)^\gamma = \frac{1}{N} \sum_{i=1}^{N} X(n, k)^\gamma.$$ 

(7)

This normalization is derived by replacing the generalized logarithmic spectra $s_\gamma(X(m, k))$ in equations (4) and (5) with $s_\gamma(X(m, k)/\hat{X}(n))$ for all $m$.

3. EVALUATION

(1) Database and Speech Analysis

The forward masking on the generalized logarithmic scale was evaluated through connected digit recognition tests. The speech data is from EDC database which consists of 35 utterances (four-connected digits) from each of 50 male and 50 female speakers. The additive noises are computer generated white noise, and the car and speech babble noises from NOISEX-92 database. The convolutional noises were simulated by the digital filters, $1 \pm 0.6 z^{-1}$. The speech and additive noise were downsampled from 16kHz to 8kHz. The speech signal was analyzed using a 20 msec Hamming window with a 5 msec frame period. A feature vector was composed of 13 DCT coefficients except the power term.

The structure of HMMs used is a left-to-right model with 16 emitting states with 4 Gaussian mixtures. The gender-independent HMMs for ten digits were trained using a total of 1600 utterances from 40 male and 40 female speakers. The recognition tests were carried out using a total of 300 test utterances from the other 10 male and 10 female speakers. The recognition performance evaluated in terms of percentage accuracy (Acc. [%] = $N - D - E^{-1} \times 100$), where $N$ tokens, $S$, $D$, and $I$ are substitution, deletion, and insertion errors, respectively.

(2) Effect of the Equal Loudness Weighting

First, we examine the effects of the generalized logarithmic scale and the equal loudness weighting without masking ($\alpha = 0$). Fig.1 shows the result of recognition tests for low frequency emphasized distortion (LFED) and for the car noise condition of -6dB SNR. Without equal loudness weighting, recognition performance degrades when $\gamma$ increases beyond 0.1. However, the generalized logarithmic scale combined with the equal loudness weighting improves recognition accuracy compared to logarithmic scale especially under the car noise conditions. Thus, the combination of the generalized logarithmic scale and the equal loudness weighting is effective for suppressing low frequency emphasized distortion.

(3) Optimization of Masking Parameters
The second experiment examines the optimal values of \( \gamma \) and subtraction coefficient \( \alpha \) with a fixed masking decay rate \( \beta \) of 0.7. Tab.1 and Fig.2 show the recognition accuracy as a function of \( \alpha \) for several values of \( \gamma \) for the convolutional noise conditions (a low frequency emphasized distortion (LFED) and a high frequency emphasized distortion (HFED)) and for the car and speech babbble noise conditions, respectively. For convolutional noises, the forward masking at \( \gamma = 0 \), which corresponds to the masking on the logarithmic scale, provides the best performance, and recognition performance slightly degrades when \( \gamma \) increases. The optimal value of \( \alpha \) is 0.8 for the values of \( \gamma = 0 \sim 0.2 \). For additive noises, the optimal value of \( \alpha \) is 0.8 to 1.2 depending on noise conditions. The forward masking at \( \gamma = 0.1 \sim 0.2 \) provides better performance than the forward masking on the logarithmic scale.

In the next experiment, the effect of the masking decay rate was examined with the fixed values of \( \gamma = 0.1, \alpha = 0.8 \). Fig.3 shows the recognition accuracy as a function of the masking decay rate \( \beta \) under the car and speech babbble noise conditions. The masking decay rate from 0.7 to 0.9 seems to be best depending on the type of noise and SNR. As a result of the above experiment, the value of \( \alpha = 0.8, \beta = 0.7, \) and \( \gamma = 0.1 \) will be used throughout the following experiments.

(4) Comparative Experiments

The recognition performance of the proposed method is compared with those of other similar conventional techniques; the forward masking on the logarithmic scale (DyMFC), the continuous spectral subtraction (CSS), and the cepstral mean subtraction (CMS). The forward masking on the mel-frequency logarithmic power spectral domain corresponds to the DyC except for the differences on the mel-frequency scale and frequency smoothing. The values of \( \gamma, \alpha, \) and \( \beta \) are set to 0.0, 0.8, and 0.7. The parameters of the CSS are as follows; the subtraction coefficient of 1.0, the flooring coefficient of 0.1, and decay rate of 0.976 which corresponds to the sum of 40 frames. The CSS is similar to the forward masking at \( \gamma = 1.0 \), but requires a larger decay rate than the DyC and the logarithmic amplitude conversion before DCT.

(a) Additive Noise Conditions

Fig.4 compare the recognition accuracies obtained by the four methods under the car noise, the white noise, and the speech babble noise conditions, respectively. The proposed method (DyMFC) outperforms the other methods at low SNR conditions. Under the white noise condition, the DyMFC improves the recognition accuracy from 78.3% for the DyMFC to 88.1% at 18dB SNR, which is also about 25% higher than that obtained by the CSS.

(b) Convolutional Noise Conditions

Table.2 compares the recognition accuracies for four methods under the low frequency emphasized distortion (LFED) and the high frequency emphasized distortion (HFED). The DyMFC attains the highest recognition ac-
The recognition accuracy even under clean condition, which is a 1.5% higher than that of the MFCC, and keeps the accuracy under both convolutional noise conditions. The recognition accuracy obtained by the proposed method (DyMFCC) are very slightly lower than those of the DyMFC, but is still higher than those of the CMS except that for the low frequency emphasized distortion.

Table 2: Comparison of the recognition accuracies by four methods under convolutional noise conditions.

<table>
<thead>
<tr>
<th>Method</th>
<th>CLEAN (%)</th>
<th>LFED (%)</th>
<th>HFED (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline/MFCC</td>
<td>87.5</td>
<td>92.1</td>
<td>90.2</td>
</tr>
<tr>
<td>DyMFCC</td>
<td>91.0</td>
<td>96.7</td>
<td>96.6</td>
</tr>
<tr>
<td>DyMFC</td>
<td>92.0</td>
<td>99.0</td>
<td>97.0</td>
</tr>
<tr>
<td>CSS</td>
<td>96.1</td>
<td>96.4</td>
<td>96.0</td>
</tr>
<tr>
<td>CMS</td>
<td>96.0</td>
<td>93.9</td>
<td>94.7</td>
</tr>
</tbody>
</table>

4. DISCUSSION

The forward masking on the generalized logarithmic scale with $\gamma = 0.1$ has been shown to improve the robustness to the additive noise over the DyMFC ($\gamma = 0.0$) while maintaining almost the same performance as that of the DyMFC for clean speech and for convolutional noise conditions. The value of $\gamma = 0.1$ is smaller than $\gamma = 0.33$ in the PLP analysis. A larger value of $\gamma$ attains further robustness to additive noise, but degrades the performance under convolutional noise. The generalized logarithmic spectra is affected by a preemphasis. In the proposed forward masking, the equal-loudness preemphasis was effective to suppress the low frequency spectral power caused by the additive noise or the low frequency emphasized distortion. Nevertheless, the low frequency part of spectra seems to be insufficiently compressed by the generalized logarithmic function. Therefore, some additional preemphasis might be needed to improve the robustness.

As mentioned previously, the proposed masking method didn't include the frequency smoothing unlike the original forward masking [3] because of its negative effect. This seems to be caused by the use of the mel-scaled filterbank spectrum in the present formulation in contrast to the linear predictive spectra on the linear frequency scale in the DyC.

5. CONCLUSION

This paper has presented the forward masking on the generalized logarithmic scale for robust speech recognition. The proposed method with $\gamma = 0.1$ in the generalized logarithmic function outperformed the conventional forward masking, the CMS, and the CSS under additive noise condition while maintaining the performance under convolutional noise.

In future work, it is necessary to examine the effect of an additional preemphasis, a flooring in masking, and frequency smoothing using a larger vocabulary recognition system.

REFERENCES