Gammatone-Domain Model Combination for Consonant Recognition in Noisy Environments

Jae Sam Yoon, Ji Hun Park, and Hong Kook Kim

Department of Information and Communications
Gwangju Institute of Science and Technology, Gwangju 500-712, Korea

{jsyoon, jh.park, hongkook}@gist.ac.kr

Abstract

In this paper, a gammatone-domain model combination method is proposed for consonant recognition in noisy environments. For this task, we first define a gammatone cepstral coefficient (GCC) as the cepstral representation of the averaged envelopes of a gammatone filtered signal. Then, we investigate a proper phonetic unit by comparing monophone, diphone, and triphone acoustic models, where it is determined from consonant recognition experiments that the diphone hidden Markov models (HMMs) provide the best performance. Next, a gammatone-domain model combination method is developed to combine the clean and noise models in the linear gammatone-envelope domain. We then evaluate the performance of the GCC-based feature and the proposed model combination on intervocalic English consonants (VCV) with 24 different consonants. It is experimentally shown that the GCC-based feature achieves a relatively higher recognition rate of 47.46% than the mel-frequency cepstral coefficients (MFCCs). Also, the model combination applied to the GCC-based diphone HMM system relatively increases the accuracy rate by 77.67% under the noisy conditions. Index Terms: consonant recognition, noise robustness, gammatone-domain, model combination, vowel-consonant-vowel (CVC) recognition

1. Introduction

Automatic consonant recognition is presently one of the most challenging topics since the performance of consonant recognition is significantly lower compared to that of vowel recognition [1][2]. Reasons for the difficulties in consonant recognition include the fact that consonants in speech signals have lower energy than vowels, tend to exhibit noise-like characteristics in the time and frequency domains [3], and have characteristics such as nasal murmur [4]. Moreover, the performance of speech recognition systems generally degrades in noisy environments because of the mismatch between noisy feature parameters and clean acoustic models [5]; consonant recognition systems are not different [6]. Thus, noise-robust feature parameters, the type of hidden Markov model (HMM) units, and noise-robust algorithms should be investigated for consonant recognition under noisy conditions.

In this paper, we basically propose a gammatone-domain model combination method to improve consonant recognition in noisy environments. It is known that gammatone-based representations provide better classification accuracy than mel-frequency cepstral coefficients (MFCCs) [7] even under noisy conditions [8][9]. That is why we select gammatone cepstral coefficients (GCCs) instead of MFCCs in this paper. In particular, a GCC is defined as the cepstral representation of the averaged envelopes of a gammatone-filtered signal. First of all, in order to select a proper phonetic unit for consonant recognition, we compare the performance of consonant recognition systems constructed by monophone, diphone, and triphone models. From consonant recognition experiments, a GCC-based diphone HMM system is chosen as a baseline system. Next, a gammatone-domain model combination method is proposed and applied to acoustic models. In other words, the gammatone-domain acoustic models are combined with noise models in the linear gammatone-envelope domain, which is similar to the conventional parallel model combination method proposed in [10]. Finally, in order to show what happens in the consonant accuracy when the noise condition is completely known, an ideal mask is calculated with a priori noise knowledge to compensate for GCC, which provides a performance bound for consonant recognition.

The outline of this paper is as follows. Section 2 describes the overview of a consonant recognition system developed in this paper. The procedure of GCC extraction and the gammatone-domain model combination method are explained in Sections 3 and 4, respectively. Next, the consonant recognition results are discussed in Section 5. Finally, background noise-aware consonant recognition for the gammatone-domain feature compensation is explained in Section 6, and this paper will be concluded in Section 7.

2. Overview of the proposed consonant recognition system

In this section, we describe the overview of a consonant recognition system developed in this paper. As shown in Fig. 1, a GCC is first extracted from the input speech signal, and then a noise model is estimated from the GCC frames classified as a...
non-speech frame using voice activity detection (VAD). This estimated noise model is used for the gammatone-domain model combination. Specifically, the estimated noise model is combined with the clean HMM in the linear gammatone-envelope domain to reduce the mismatch between the GCC and clean HMM. A more detailed description of the gammatone-domain model combination method is explained in Section 4.

3. Gammatone-domain feature extraction
At first, we describe the procedure of obtaining GCC feature from input speech. Fig. 2 shows a schematic diagram for extracting GCC parameters. Input signals with a sampling rate of 25 kHz are first decomposed into auditory spectral signals by employing gammatone filterbanks [11] with 64 channels, in which the center frequencies are linearly spaced on an ERB-scale [12] from 50 Hz to 12.5 kHz. Here, the auditory spectral signals are windowed using a rectangular window with a time resolution of 25 ms and a frame rate of 100 Hz. The envelopes for the auditory spectral signals $env(i, j)$ for the $i$-th gammatone filterbank channel and the $j$-th frame are then computed as follows.

$$
env(i, j) = \left[ \sum_{n=0}^{N-1} x^{i,j}(n) \right] (1)
$$

where $x^{i,j}(n)$ represents the $n$-th auditory spectral signal for the $i$-th gammatone filterbank channel and the $j$-th frame, and $N$ is the number of speech samples within a frame. In this paper, $N$ is set to 625, which corresponds to 25 ms at a sampling rate of 25 kHz.

GCCs are extracted by applying a discrete cosine transform (DCT) to the envelopes, $env(i, j)$. That is, the conversion equation from the envelopes to GCCs is defined as

$$
c(j,k) = \sum_{i=1}^{I} \log(env(i,j)) \cdot \cos\left(\frac{k\pi}{I}(i-0.5)\right) (2)
$$

where $I (= 64)$ is the total number of the gammatone filterbanks and $k$ indicates the index number of GCCs ranging from 0 to 12.

4. Gammatone-domain model combination
A gammatone-domain model combination method is developed to obtain a noise-robust model without a priori noise information. In the model combination method proposed in [10] the previously trained clean model is combined with the noise model in the linear spectral domain. Thus, the clean and noise model in the cepstral domain should be transformed in the linear spectral domain. Similarly, the clean and noise models in the GCC domain are first transformed into the linear gammatone-envelope domain defined in (1). And then, two models are combined in the linear gammatone-envelope domain, and transformed back to the GCC domain. Of course, the noise model is estimated from the GCC domain by using non-speech frames classified using a VAD algorithm. In this paper, we utilize the VAD algorithm in the extended version of an ETSI front-end [13].

Fig. 3 shows a procedure of combining clean and noise models for GCCs. First of all, the mean and variance of the clean model in the GCC domain is transformed by taking an inverse discrete cosine transform (IDCT) into the log gammatone-envelope domain as

$$
\hat{\mu}^i = C^{-1}\mu^c \\
\hat{\Sigma}^i = C^{-1}\Sigma^c (C^{-1})^T (3)
$$

where $\mu^c$ and $\Sigma^c$ are the mean and variance of the clean model in the GCC domain, respectively. In addition, $\hat{\mu}^i$ and $\hat{\Sigma}^i$ are the mean and variance of the clean model in the log gammatone-envelope domain. Here, $C$ is the matrix representing the DCT with $C_{lk} = \cos(k\pi(l-0.5)/I)$, where $i$ and $k$ are the indices of the gammatone filterbank and cepstrum, respectively, and $I (= 64)$ is the number of the gammatone filterbanks. Similarly, the noise model parameters, $\tilde{\mu}^c$ and $\tilde{\Sigma}^c$, are transformed to the log gammatone-envelope domain $\tilde{\mu}^i$ and $\tilde{\Sigma}^i$.

Next, the clean and noise model parameters in the log gammatone-envelope domain are converted in the linear gammatone-envelope domain by the equation of

$$
\mu_i = \exp(\mu_i^i + \Sigma_i^i/2) (5) \\
\Sigma_{im} = \mu_i \mu_m \left[\exp(\Sigma_{im}^i) - 1\right] (6) \\
\tilde{\mu}_i = \exp(\tilde{\mu}_{i} + \tilde{\Sigma}_{i}^i/2) (7) \\
\tilde{\Sigma}_{im} = \tilde{\mu}_i \tilde{\mu}_m \left[\exp(\tilde{\Sigma}_{im}^i) - 1\right] (8)
$$

where $\mu$ and $\Sigma$ are the mean and variance of clean model in the linear gammatone-envelope domain, respectively, $\tilde{\mu}$ and $\tilde{\Sigma}$ are the mean and variance of noise model in the linear gammatone-envelope domain, $\mu_i$ is the $i$-th element of a vector $\mu$, and $\Sigma_{im}$ is the $im$-th element of a matrix $\Sigma$.

In the linear gammatone-envelope domain, the clean and noise models are combined as

$$
\hat{\mu}_i = \mu_i + \tilde{\mu}_i (9) \\
\hat{\Sigma}_{im} = \Sigma_{im} + \tilde{\Sigma}_{im} (10)
$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the mean and variance of the estimated noisy models in the linear gammatone-envelope domain. After
that, the noisy model is converted back into the log gammatone-envelope domain.

\[
\hat{\mu}_i^l = \log(\hat{\mu}_i) - \frac{1}{2} \log \left( \hat{\Sigma}_{ii}/\hat{\mu}_i^2 + 1 \right) \tag{11}
\]

\[
\hat{\Sigma}_{im} = \log \left( \hat{\Sigma}_{im}/\hat{\mu}_i\hat{\mu}_m + 1 \right) \tag{12}
\]

Finally, the combined noisy model in the GCC domain is obtained by applying a DCT to (11) and (12).

### 5. Consonant recognition experiments

In this section, we try to find a phonetic unit by comparing the performance of GCC-based HMM systems employing different units such as monophone, diphone, and triphone HMMs. And then, the performance of the gammatone-domain model combination described in Section 4 is evaluated.

#### 5.1. Task and database

The corpus used to train and test acoustic models is the inter-vocalic English consonants (VCV), where nine possible combinations of three vowels /i/, /u/, and /æ/ are combined with 24 consonants. The training set consists of 6,667 tokens recorded by 8 male and 8 female speakers. To simulate a noisy environment, seven test sets are built from 4 male and 4 female speakers; one clean condition, and six noise conditions with different noise types. Each test set consists of 384 tokens, where each talker says the 24 consonants twice. The test sets are provided as two types: single-channel noise sets and dual-channel noise sets. In the single-channel noise sets, clean speech and additive noise signals are recorded in the form of a mono signal. In the dual-channel noise sets, clean speech and noise signals are separately recorded as a stereo signal; for further details, refer to [14].

#### 5.2. HMM training

In order to construct a noise-robust baseline system, various scenarios of acoustic models are trained and tested by using HTK v.3.2 [15]. First, GCC-based and MFCC-based monophone HMMs are compared to show the advantage of GCCs over MFCCs in a noisy environment. For the GCC feature, the GCC-based 32-mixture monophone HMMs are trained by using a 39-dimensional GCC feature vector composed of 13 static GCC concatenated with their deltas and delta-deltas. On the other hands, the MFCC-based 24-mixture monophone HMMs are trained for the comparison. In addition, GCC-based 26-mixture diphone HMMs and 24-mixture triphone HMMs are trained, where the number of mixtures is experimentally selected. As a result, the numbers of models are 151 and 367 for diphone HMMs and triphone HMMs, respectively.

#### 5.3. Phonetic unit selection

The phonetic unit for the baseline system is determined based on the consonant recognition experiments using the single-channel noise sets. Table 1 shows the performance of consonant recognition according to the features and phonetic units used in the system; MFCC and Ratemap in Table 1 are the performance of MFCC-based and ratemap-based HMM systems constructed according to [14]. As shown in Table 1, the consonant accuracy of Ratemap is lower than that of MFCC. However, the consonant accuracy of the GCC-based monophone HMMs is relatively higher by 12.00% than that of the MFCC-based monophone HMMs under noisy conditions. This difference suggests that GCC is better than MFCC for consonant recognition in noisy environments. Moreover, the consonant accuracy of the GCC-based diphone HMMs is relatively higher by 47.46% than that of the GCC-based monophone HMMs. Based on these experimental results, the GCC-based diphone HMMs are finally selected as a baseline system in this paper.

#### 5.4. Gammatone-domain model combination

Table 2 shows the comparison of consonant accuracies of GCC and the proposed model combination applied to GCC. Here, the performance of GCC refers to as the performance of the GCC-based diphone HMM system constructed in Section 5.2. In the clean condition, the performance of GCC+GMC degrades more than that of the GCC-based diphone HMM system due to the inexact noise model estimation from the clean speech signal. However, in noisy conditions, the consonant accuracy of GCC+GMC is relatively increased by 77.67% compared with

---

**Table 1:** Comparison of consonant recognition accuracies (%) on single-channel noise sets according to different types of features and phonetic units.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Phonetic Unit</th>
<th>Clean Condition</th>
<th>Noise Condition</th>
<th>Relative Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Testset 1</td>
<td>Testset 2</td>
<td>Testset 3</td>
<td>Testset 4</td>
</tr>
<tr>
<td>MFCC</td>
<td>Monophone</td>
<td>88.54</td>
<td>11.98</td>
<td>5.47</td>
</tr>
<tr>
<td></td>
<td>Ratemap</td>
<td>81.31</td>
<td>5.47</td>
<td>3.12</td>
</tr>
<tr>
<td>GCC</td>
<td>Monophone</td>
<td>86.98</td>
<td>12.50</td>
<td>5.47</td>
</tr>
<tr>
<td></td>
<td>Diphone</td>
<td>89.32</td>
<td>14.58</td>
<td>13.80</td>
</tr>
<tr>
<td></td>
<td>Triphone</td>
<td>77.34</td>
<td>10.94</td>
<td>11.72</td>
</tr>
</tbody>
</table>

**Table 2:** Consonant recognition accuracies (%) of the proposed gammatone-domain model combination on single-channel noise sets.

<table>
<thead>
<tr>
<th>Model Combination</th>
<th>Clean Condition</th>
<th>Noise Condition</th>
<th>Relative Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Testset 1</td>
<td>Testset 2</td>
<td>Testset 3</td>
</tr>
<tr>
<td>GCC (Diphone)</td>
<td>89.32</td>
<td>14.58</td>
<td>13.80</td>
</tr>
<tr>
<td>GCC+GMC</td>
<td>82.29</td>
<td>14.84</td>
<td>23.18</td>
</tr>
</tbody>
</table>
that of the GCC-based diphone HMMs.

6. Background noise-aware consonant recognition

In this section, we attempt to evaluate the performance of a consonant recognition system if we can exactly know the background noise added to clean speech. For this task, we use the dual-channel noise sets of the consonant database, and compute the ideal mask defined as [16]

\[
m_{\text{ideal}}(i, j) = \frac{e_{\text{env}}(i, j)}{e_{\text{env}}(i, j) + e_{\text{env}}(i, j)}
\]

where \(e_{\text{env}}(i, j)\) and \(e_{\text{env}}(i, j)\) are the averaged envelopes of the speech and noise signals for the \(i\)-th frequency channel and \(j\)-th frame, respectively, as described in (1). Next, the proposed gammatone-domain feature compensation method is applied to the background noise-aware consonant recognition. In this method, the noise is reduced using the ideal mask in the linear gammatone-envelope domain as follows.

\[
\tilde{e}_{\text{env}}(i, j) = m_{\text{ideal}}(i, j) \cdot e_{\text{env}}(i, j)
\]

Finally, the noise-compensated GCC can be obtained from \(\tilde{e}_{\text{env}}(i, j)\) using (2). The consonant accuracies of the gammatone-domain feature compensation method (GCC+Ideal Mask) are shown in Table 3. Note that GCC is the performance of the GCC-based diphone HMM system described in Section 5.2. As shown in Table 3, the performance bound is outstandingly higher than the GCC-based diphone HMM system under all the noise conditions.

7. Conclusion

In order to improve the performance of consonant recognition in noisy environments, we first defined a gammatone cepstral coefficient (GCC) and constructed a GCC-based diphone HMM system. Next, the proposed gammatone-domain model combination (GMC) method was applied to the GCC-based diphone HMM system. In the GMC method, clean models were combined with an estimated noise model in the gammatone envelope domain. In addition, we evaluated the performance of a consonant recognition system to show the performance bound when noise condition was exactly known prior. Subsequently, it was shown from experimental results on a task of intervocalic English consonants (VCV) that the gammatone-domain model combination relatively increased the recognition rate by 77.67\% compared to that of the GCC-based diphone HMM system under the noisy conditions.

8. Acknowledgements

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-314-D00245) and partially by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the Institute of Information Technology Advancement (IITA-2008-C1090-0801-0017).

Table 3: Consonant accuracies (%) of background noise-aware consonant recognition for dual-channel noise sets.

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Clean Condition</th>
<th>Noise Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Testset 1</td>
<td>Testset 2</td>
</tr>
<tr>
<td>GCC (Diphone)</td>
<td>89.32</td>
<td>14.58</td>
</tr>
<tr>
<td>GCC+Ideal Mask</td>
<td>89.32</td>
<td>46.09</td>
</tr>
</tbody>
</table>

9. References