Canonicalization of Feature Parameters for Automatic Speech Recognition

Takashi Fukuda and Tsuneo Nitta

Graduate School of Engineering, Toyohashi University of Technology, Japan
fukuda@vox.tutkie.tut.ac.jp, nitta@tutkie.tut.ac.jp

Abstract

Acoustic models (AMs) of an HMM-based classifier include various types of hidden variables such as gender type, speaking rate, and acoustic environment. If there exists a canonicalization process that reduces the influence of the hidden variables from the AMs, a robust automatic speech recognition (ASR) system can be realized. In this paper, we describe the configuration of a canonicalization process targeting gender type as a hidden variable. The proposed canonicalization process is composed of multiple distinctive phonetic feature (DPF) extractors corresponding to the hidden variable and a DPF selector in which the distance between input DPF and AMs is compared. In a DPF extraction stage, an input sequence of acoustic feature vectors is mapped onto three DPF spaces corresponding to male, female, and neutral voice by using three multilayer neural networks (MLNs). Experiments are carried out by comparing (A) the combination of the canonicalized DPF and a single HMM classifier, and (B) the combination of a single acoustic feature (MFCC) and multiple HMM classifiers. The result shows that the proposed canonicalization method outperforms both of the conventional ASR with MFCC and a single HMM and the ASR with multiple HMMs in spite of less memories and computation time.

1. Introduction

Recently, a current automatic speech recognition (ASR) system can achieve recognition accuracy more than 90% when the speech is uttered clearly in low-noise environments. This is mainly due to the powerful support of a stochastic classifier, or an HMM-based classifier, and large scale speech corpora, however, because acoustic models (AMs) of the HMM-based classifier include hidden variables such as gender type, speaking rate, and acoustical environment, a single HMM classifier can not obtain enough performance every time. Several strategies of estimating the AMs using speech data collected in various conditions have been investigated toward this problem [1]. These strategies can represent various types of acoustic varieties in the AMs.

On the other hand, an approach of decoding in parallel with multiple HMMs corresponding to hidden variables has recently been proposed ([2, 3] see Figure 1). Multi-path acoustic modeling, that represents hidden variables with several paths in the same AM instead of applying multiple HMMs, was also proposed [4, 5].

These approaches may achieve robust ASR systems in various aspects, however, they need large amount of memories and computation time. If there exists a canonicalization process of feature parameters that reduces the influence of the hidden variables from the AMs, the robust ASR system can be realized in low cost.

In this paper, we describe the configuration of the canonicalization process targeting gender type as the hidden variable. In our proposed method, three distinctive phonetic feature (DPF) extractors composed of multilayer neural networks (MLNs) map an input sequence of acoustic feature vectors onto three DPF spaces corresponding to male, female, and neutral voice (see Figure 2). Then, a DPF selector, in which the distance between input DPF and AMs is compared, outputs a nearest DPF to AMs as a canonicalized DPF. Canonicalization described above is realized by introducing a DPF space between an acoustic feature space and the AM of an HMM classifier for the first time.
This paper is organized as follows. Section 2 outlines implementation of a DPF extractor, and Section 3 explains the canonicalization process of feature parameters. Section 4 describes an experimental setup and results, and provides a discussion. Finally, Section 5 finishes with some conclusions.

2. Overview of a DPF extractor

This section describes an overview of a DPF extractor which plays an important role in a canonicalization process of feature parameters.

The configuration of the DPF extractor is illustrated in Figure 3. At an acoustic feature extraction stage, firstly, input speech is converted into local features (LFs) that represent a variance of spectrum along time and frequency axes. LFs are then entered into an MLN with four layers, including two hidden layers, after combining a current frame with the other two frames that are 3-points before and after the current frame. The MLN has 45 output units (15 dim. × 3 fr.) with 15 dimensions each. The two hidden layers consist of 256 and 96 units from the input layer. A DPF set is created by using a backpropagation algorithm to adjust the distances among phonemes by using the DPF vectors and initial training of AMs.

Next, AMs are redesigned using the three DPF extractors and the DPF selector. The DPF selector outputs DPF vectors, that consist of three DPF vectors (a preceding context DPF vector, a current DPF vector, and a following context DPF vector) with 15 dimensions. The interval in which a value of one DPF element “mora” is bigger than 0.5 is recognized as a vowel interval. The DPF elements in the vowel intervals have higher reliability than that in consonantal parts. When DM between input candidate vectors of DPF male and the AMs is close (within ±25%) to DM between another input candidate vectors of DPF female and the AMs, the DPF selector extracts the other input candidate vectors DPF neutral as the canonicalized DPF. Finally, the canonicalized DPF vectors are input to the HMM classifier. These iterative training, or

3. Canonicalization of feature parameters

In this section, we describe the configuration of a canonicalization process of feature parameters. In the canonicalization process shown in Figure 2, three MLNs are firstly used to map an input sequence of acoustic feature vectors onto three DPF spaces corresponding to male, female and neutral voice (three DPFs output from each MLN are called DPF male, DPF female, and DPF neutral, respectively). Two DPF extractors for male and female voice are trained independently with male and female data in the D1 data set in Section 4.1. The DPF extractor for neutral voice is designed with both male and female data in D1. The distortion of DPF vectors corresponding to male, female, or neutral voice is less than that of DPF vectors extracted from a single MLN. Thus, it is expected to reduce the influence of the hidden variable by using the DPF vectors corresponding to speaker’s characteristics as the input of an HMM classifier. Next, the canonicalization process needs a DPF selector described below.

To select the desirable DPF vectors that correspond to the gender type of input speech, there are several strategies including a distortion measure and an information criterion. In the followings, firstly, we explain about labeled DPF vectors and initial training of AMs.

(1) Two MLNs corresponding to male and female voice are trained independently using male and female speech in D1 data set described in Section 4.1.

(2) Input speech is mapped onto DPF spaces according to the gender type by using the above MLN. Here, MLN is selected using label information in D1 data set.

(3) AMs are trained using DPF vectors selected along the above mentioned process.

Next, AMs are redesigned using the three DPF extractors and the DPF selector. The DPF selector outputs DPF vectors, that minimize the distance \( d \) in the following equation, as the canonicalized DPF every words.

\[
D_M(p_j, q_j) = \left| \mu_j - \mu_j \right| \Sigma_j^{-1} \left( \log(p_j) - \mu_j \right)
\]

where \( p_j \) is the input DPF vector of \( j \)-th frame, \( q_j \) is the distribution parameter set in the middle state of \( j \)-th HMM, and \( D_M \) is Mahalanobis distance between \( p_j \) and \( q_j \). \( N \) and \( M \) are total number of frames in a vowel interval and number of monophone models in a word, respectively. \( \mu \) and \( \Sigma \) are mean vector and covariance matrix in the model. Here, \( D_M \) is calculated between the HMM (mix = 1) and the input DPF vectors in vowel intervals. The interval in which a value of one DPF element “mora” is bigger than 0.5 is recognized as a vowel interval. The DPF elements in the vowel intervals have higher reliability than that in consonantal parts. When \( D_M \) between input candidate vectors of DPF male and the AMs is close (within ±25%) to \( D_M \) between another input candidate vectors of DPF female and the AMs, the DPF selector extracts the other input candidate vectors DPF neutral as the canonicalized DPF. Finally, the canonicalized DPF vectors are input to the HMM classifier.
4. Experiments

4.1. Speech database

The following two data sets were used:

**D1.** Acoustic model design set with clean speech:
A subset of “ASJ (Acoustic Society of Japan) Continuous Speech Database”, consisting of 9,003 sentences uttered by 30 male and 30 female speakers, respectively (16 kHz, 16-bit)[10].

**D2.** Test data set with clean speech:
A subset of “Tohoku University and Matsushita Spoken Word Database”, consisting of 100 words uttered by 10 unknown male and female speakers each[11]. Total number of words is 1,962. The sampling rate was converted from 24 kHz to 16 kHz.

4.2. Experimental Setup

An input speech is sampled at 16 kHz and a 512-point FFT of the 25 ms Hamming-windowed speech segment is applied every 10 ms. The resultant FFT power spectrum is then integrated into 24-ch BPFs output with mel-scaled center frequencies. At the acoustic feature extraction stage, an output of BPF bank are converted into local features (LFs), and then LFs are mapped to DPFs.

The D1 data set was used to design 43 Japanese monophone HMMs with five states and three loops. In the HMM, output probabilities are represented in normal distribution (ND) for MFCC parameter set and logarithmic normal distribution (LND) with only negative skewness for DPFs[9], and diagonal matrices are used.

4.3. Experimental Results

Speaker-independent isolated spoken-word recognition tests were carried out with the D2 data set. In the experiments, four types of ASR systems illustrated in Figure 4 were investigated. In the baseline 1, the input of HMM is the conventional acoustic feature set with 38 dimensions which consists of MFCC with CMN, dynamic features ($\Delta t$, $\Delta ^2 t$), $\Delta P$ and $\Delta \Delta P$.

Parallel HMMs of baseline 2 in Figure 4 (b) are followed by hypothesis selector that recognizes an HMM with a targeting hidden variable owner based on a maximum likelihood (ML) criterion. An original DPF in Figure 4 (c) indicates a method of using a single DPF extractor which is designed both male and female voice in D1 data set.

Figure 5 shows the experimental results. The canonicalized DPF outperformed the both baseline systems except the case of a single mixture. The original DPF with a single DPF extractor also yielded better performance than the baseline systems. This gain in performance is considered to be as follows. The DPF vectors are originally specified for classifying phonemes even if the hidden variable of gender is included, and the canonicalized DPF vectors can form a feature space that is independent of the targeting hidden variable of the gender type.

4.4. Discussion

The proposed canonicalization process has three types of DPF extractors (two DPF extractors for male and female voice, and one DPF extractor for neutral voice). Here, an effect of adding the DPF extractor for neutral voice is investigated. Figure 6 shows the recognition result when eliminating the DPF extractor for neutral voice from the canonicalization process. Tables 1 and 2 provide the selection rate to three DPFs
DPFs and the single HMM showed better performance in the small computation time and memories than the conventional ASR system with multiple HMMs. The proposed canonicalization process can reduce the influence of gender type as hidden variable from the AMs.

In feature work, we will discuss a canonicalization method targeting the other types of hidden variables such as speaking style and acoustic environment.

Acknowledgements
This work was supported in The 21st Century COE Program “Intelligent Human Sensing”, from the ministry of Education, Culture, Sports, Science and Technology. And the authors want to thank the Hori Information Science Promotion Foundation for supporting this work.

References

5. Conclusion

The canonicalization process of the feature parameters was proposed. The canonicalization is realized by introducing DPF space between acoustic feature space and the AM of HMM classifier. In an experiment using an isolated-spoken word recognition task, the combination of the canonicalized DPFs and the single HMM showed better performance in the small computation time and memories than the conventional ASR system with multiple HMMs. The proposed canonicalization process can reduce the influence of gender type as hidden variable from the AMs.

When the DPF extractor for neutral voice was eliminated from canonicalization process, the recognition performance is reduced as shown in Figure 6. On the other hands, in Table 2, DPFs of around 13% in D2 data set (Tohoku Univ. corpus) does not correspond to input gender type. These selection errors can be collected into the new category of the neutral voice (see Table 1), thus the DPF extractor for neutral voice contributes to the improvement of performance.

Table 1. The DPF selection rate (DSR) [%] and its gender type accuracy (ACC) [%] with a DPF extractor for neutral voice.

<table>
<thead>
<tr>
<th></th>
<th>ASJ corpus (Training set)</th>
<th>Tohoku Univ. corpus (Test set)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSR</td>
<td>ACC</td>
</tr>
<tr>
<td>DPFmale</td>
<td>34.7</td>
<td>100.0</td>
</tr>
<tr>
<td>DPFfemale</td>
<td>31.5</td>
<td>99.9</td>
</tr>
<tr>
<td>DPFneutral</td>
<td>33.8</td>
<td>39.1</td>
</tr>
</tbody>
</table>

(DPF_{male}, DPF_{female}, and DPF_{neutral}) and its accuracy of gender type.

When the DPF extractor for neutral voice was eliminated from canonicalization process, the recognition performance is reduced as shown in Figure 6. On the other hands, in Table 2, DPFs of around 13% in D2 data set (Tohoku Univ. corpus) does not correspond to input gender type. These selection errors can be collected into the new category of the neutral voice (see Table 1), thus the DPF extractor for neutral voice contributes to the improvement of performance.

Table 2. The DPF selection rate (DSR) [%] and its gender type accuracy (ACC) [%] without a DPF extractor for neutral voice.

<table>
<thead>
<tr>
<th></th>
<th>ASJ corpus (Training set)</th>
<th>Tohoku Univ. corpus (Test set)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSR</td>
<td>ACC</td>
</tr>
<tr>
<td>DPFmale</td>
<td>48.5</td>
<td>99.5</td>
</tr>
<tr>
<td>DPFfemale</td>
<td>51.5</td>
<td>96.6</td>
</tr>
</tbody>
</table>

(DPF_{male}, DPF_{female}, and DPF_{neutral}) and its accuracy of gender type.

When the DPF extractor for neutral voice was eliminated from canonicalization process, the recognition performance is reduced as shown in Figure 6. On the other hands, in Table 2, DPFs of around 13% in D2 data set (Tohoku Univ. corpus) does not correspond to input gender type. These selection errors can be collected into the new category of the neutral voice (see Table 1), thus the DPF extractor for neutral voice contributes to the improvement of performance.