Designing Multiple Distinctive Phonetic Feature Extractors for Canonicalization by Using Clustering Technique

Takashi Fukuda, Muhammad Ghulam and Tsuneo Nitta

Graduate School of Engineering, Toyohashi University of Technology, Japan
(Fukuda, ghulam)@vox.tutkiet.tut.ac.jp, nitta@tutkiet.tut.ac.jp

Abstract

Acoustic models of an HMM-based classifier include various types of hidden factors such as speaker-specific characteristics and acoustic environments. If there exist a canonicalization process that represses the decrease in acoustic-likelihood among categories resulted from hidden factors, a robust ASR system can be realized. We have previously proposed the canonicalization process of feature-parameters composed of three distinctive phonetic feature (DPF) extractors focused on a gender factor. This paper describes an attempt to design multiple DPF extractors corresponding to unspecified hidden factors, as well as to introduce a noise suppressor that is targeted for the canonicalization of a noise factor. In an experiment on Japanese version AURORA2 database (AURORA2-J), the proposed system achieved significant improvements when combining the canonicalization process with the noise reduction technique based on a two-stage Wiener filter.

1. Introduction

Many approaches have been investigated to realize a robust automatic speech recognition (ASR) system, however, the ASR system that shows enough performance anytime and everywhere could not be realized now. One of the reasons is that the acoustic models (AMs) of an HMM-based classifier include many hidden factors such as speaker-specific characteristics that includes gender types and speaking styles, and acoustic environments including channel characteristics and ambient noise. To overcome these difficulties, an approach of decoding in parallel with multiple HMMs corresponding to hidden factors has recently been proposed ([1], [2], see Fig. 1). Multi-path acoustic modeling, that represents hidden factors with several paths in the same AM instead of applying multiple HMMs, was also proposed [3].

In contrast, we have previously proposed a feature-parameter canonicalization process composed of multiple distinctive phonetic feature (DPF) extractors (see Fig. 2). Here, the canonicalization of feature-parameters is defined as a process to repress the decrease in acoustic-likelihood differences among categories resulted from the above-mentioned hidden factors. In our previous work [4], the canonicalization process was composed of three DPFs which are explicitly designed by assigning a data set to each DPF. Namely, two DPF extractors for male and female voice are trained independently with a male and a female data set, respectively, and the other DPF extractor for neutral voice is designed with both male and female data sets. As a result, the canonicalization process could form a feature space that is independent of the gender factor and improve the performance on an experiment using an isolate spoken word recognition task. However, there exist many other factors, and we can not design all the DPF extractors that correspond with them. Moreover, the canonicalization should be realized by not only DPF extractors, but also by other appropriate mechanisms according to targeting factors.

This paper describes an attempt to design multiple DPF extractors corresponding to hidden factors, as well as to introduce a noise suppressor that is expected to canonicalize a noise factor. In the DPF extractor design, we apply a sophisticated clustering technique to cope with multiple factors. At the same time, a DPF selector in Fig. 2 which selects a desirable DPF as a canonicalized DPF is also modified to isolate its design from HMM classifier design, while the previous DPF selector design is coupled with the classifier design. The isolation is an important idea for a distributed speech recognition (DSR). On the other hand, in the noise suppressor design, a noise reduction technique based on a two-stage Wiener filter proposed by ETSI is applied to the canonicalization process as a preprocessing [5]. Experiments are carried out using the Japanese version of AURORA2 database (AURORA2-J [6]).

Fig. 1 A single feature extractor and multiple HMM classifiers.

Fig. 2 Multiple DPF extractors and a single HMM classifier.
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. The configuration of the DPF extractor[4] 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[7]. LFs are then entered into an MLN with four layers, including two hidden layers, after combining a current frame, \(x_t\), with the other two frames that are 3-points before and after the current frame \(x_{t-3}, x_{t+3}\). The MLN has 45 output units \((15 \times 3)\) corresponding to a set of triphones, or context-dependent DPFs, that consist of three DPF vectors (a preceding context DPF vector, a current DPF vector, and a following context DPF vector) with 15 dimensions each. The two hidden layers consist of 256 and 96 units from the input layer. Fifteen DPF elements of mora, high, low, nil (an intermediate expression of “high / low”), anterior, back, nil (an intermediate expression of “anterior / back”), coronal, plosive, affricative, continuant, voiced, unvoiced, nasal, and semi-vowel are used to catch the balance of phoneme configuration in a DPF space[8]. The MLN is trained by using a backpropagation algorithm to output a value of 1 for the corresponding DPF elements with an input phoneme and its adjacent phonemes (a set of triphone). In this paper, the MLN output with 45 dimensions was orthogonalized and the resultant vector with 33 dimensions was used in the HMM classifier[8].

3. Canonicalization process

3.1. Configuration of a canonicalization process

Fig. 4 shows the block diagram of the proposed canonicalization process. In this process, an input speech is firstly de-noised by using a 2-stage Wiener filter of the ETSI standard Advanced DSR front-end ES202 (WI008) [5]. The noise suppression used in WI008 is very powerful, however, gains in ASR performance depend on speakers[9], and a succeeding feature extraction stage is expected to enhance such speech. Next, the de-noised speech is processed at a BPF bank and an LF extractor, and then input into N DPF extractors. Each DPF extractor is designed by using a clustering algorithm described in the next section. The output \(\{X_1(t), X_2(t), \ldots, X_N(t)\}\) of the DPF extractors are
Table 1 Word accuracy [%] of baseline (WI008)

<table>
<thead>
<tr>
<th>Clean</th>
<th>Subway</th>
<th>Babble</th>
<th>Car</th>
<th>Exhibition</th>
<th>Average</th>
<th>Restaurant</th>
<th>Street</th>
<th>Airport</th>
<th>Station</th>
<th>Average</th>
<th>Subway M</th>
<th>Street M</th>
<th>Average</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>98.43</td>
<td>98.37</td>
<td>98.57</td>
<td>98.40</td>
<td>98.54</td>
<td>98.43</td>
<td>98.57</td>
<td>98.40</td>
<td>98.57</td>
<td>98.54</td>
<td>98.62</td>
<td>98.64</td>
<td>98.63</td>
<td>98.68</td>
</tr>
<tr>
<td>20 dB</td>
<td>97.81</td>
<td>98.43</td>
<td>98.69</td>
<td>97.72</td>
<td>98.11</td>
<td>96.64</td>
<td>96.43</td>
<td>97.94</td>
<td>98.03</td>
<td>97.11</td>
<td>97.64</td>
<td>96.67</td>
<td>97.16</td>
<td>97.52</td>
</tr>
<tr>
<td>15 dB</td>
<td>94.17</td>
<td>96.07</td>
<td>97.94</td>
<td>95.80</td>
<td>96.00</td>
<td>92.17</td>
<td>94.23</td>
<td>96.00</td>
<td>96.14</td>
<td>93.80</td>
<td>93.65</td>
<td>93.73</td>
<td>95.00</td>
<td></td>
</tr>
<tr>
<td>10 dB</td>
<td>86.18</td>
<td>89.18</td>
<td>95.17</td>
<td>90.34</td>
<td>90.22</td>
<td>81.46</td>
<td>88.27</td>
<td>89.02</td>
<td>89.36</td>
<td>87.03</td>
<td>85.69</td>
<td>86.91</td>
<td>88.30</td>
<td></td>
</tr>
<tr>
<td>5 dB</td>
<td>66.51</td>
<td>69.26</td>
<td>83.21</td>
<td>73.19</td>
<td>73.00</td>
<td>59.90</td>
<td>72.46</td>
<td>73.22</td>
<td>77.32</td>
<td>70.84</td>
<td>62.57</td>
<td>68.83</td>
<td>69.07</td>
<td>68.66</td>
</tr>
<tr>
<td>0 dB</td>
<td>36.97</td>
<td>31.92</td>
<td>48.05</td>
<td>37.70</td>
<td>38.60</td>
<td>23.43</td>
<td>44.81</td>
<td>49.08</td>
<td>49.24</td>
<td>39.42</td>
<td>52.70</td>
<td>40.24</td>
<td>38.52</td>
<td>38.66</td>
</tr>
<tr>
<td>Average</td>
<td>76.29</td>
<td>76.96</td>
<td>84.61</td>
<td>78.95</td>
<td>79.20</td>
<td>70.60</td>
<td>79.08</td>
<td>79.43</td>
<td>82.12</td>
<td>77.81</td>
<td>74.48</td>
<td>77.26</td>
<td>75.87</td>
<td>77.98</td>
</tr>
</tbody>
</table>

Table 2 Word accuracy [%] of the proposed canonicalized DPF (No. of clusters : N = 4)

<table>
<thead>
<tr>
<th>Clean</th>
<th>Subway</th>
<th>Babble</th>
<th>Car</th>
<th>Exhibition</th>
<th>Average</th>
<th>Restaurant</th>
<th>Street</th>
<th>Airport</th>
<th>Station</th>
<th>Average</th>
<th>Subway M</th>
<th>Street M</th>
<th>Average</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 dB</td>
<td>98.31</td>
<td>97.57</td>
<td>99.01</td>
<td>97.93</td>
<td>98.21</td>
<td>90.07</td>
<td>97.91</td>
<td>96.11</td>
<td>98.29</td>
<td>95.82</td>
<td>98.59</td>
<td>97.91</td>
<td>98.25</td>
<td>97.26</td>
</tr>
<tr>
<td>15 dB</td>
<td>95.79</td>
<td>96.12</td>
<td>98.50</td>
<td>96.04</td>
<td>96.61</td>
<td>88.97</td>
<td>95.77</td>
<td>95.87</td>
<td>96.03</td>
<td>94.16</td>
<td>95.85</td>
<td>96.22</td>
<td>96.04</td>
<td>95.58</td>
</tr>
<tr>
<td>10 dB</td>
<td>89.21</td>
<td>90.15</td>
<td>96.27</td>
<td>91.99</td>
<td>91.91</td>
<td>83.17</td>
<td>89.01</td>
<td>89.91</td>
<td>89.03</td>
<td>88.03</td>
<td>87.51</td>
<td>90.08</td>
<td>88.81</td>
<td>89.14</td>
</tr>
<tr>
<td>5 dB</td>
<td>67.52</td>
<td>69.89</td>
<td>77.50</td>
<td>73.24</td>
<td>72.54</td>
<td>66.74</td>
<td>75.80</td>
<td>71.34</td>
<td>79.20</td>
<td>72.02</td>
<td>64.25</td>
<td>69.74</td>
<td>67.00</td>
<td>71.02</td>
</tr>
<tr>
<td>0 dB</td>
<td>38.97</td>
<td>36.76</td>
<td>50.27</td>
<td>38.85</td>
<td>41.46</td>
<td>37.53</td>
<td>40.11</td>
<td>41.52</td>
<td>51.82</td>
<td>42.75</td>
<td>37.53</td>
<td>39.03</td>
<td>36.39</td>
<td>40.95</td>
</tr>
<tr>
<td>-5 dB</td>
<td>15.58</td>
<td>15.05</td>
<td>15.33</td>
<td>14.93</td>
<td>15.17</td>
<td>15.79</td>
<td>16.84</td>
<td>16.69</td>
<td>16.98</td>
<td>18.34</td>
<td>17.20</td>
<td>17.77</td>
<td>16.82</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>77.96</td>
<td>78.10</td>
<td>84.31</td>
<td>80.21</td>
<td>80.14</td>
<td>73.48</td>
<td>78.72</td>
<td>78.97</td>
<td>83.05</td>
<td>78.56</td>
<td>79.99</td>
<td>78.60</td>
<td>77.30</td>
<td>78.94</td>
</tr>
</tbody>
</table>

Finally, input to a DPF selector and a desired canonicalized DPF vector $X_p(t)$ is selected. In the DPF selector, the vector of $n$-th DPF extractor which satisfies the following equation is output to the HMM classifier.

$$n = \arg \min_{p=1,...,P} \left\{ \sum_{k=1}^{K} \min_{\chi} ||R - X_p(t)||^2 \right\}$$

(1)

Here, $\chi$, $R$, and $T$ are a set of DPF vectors, a DPF vector of a phoneme, and the number of frames in an utterance, respectively. In the case of a phoneme $/\alpha/$, $R$ is $(1,0,1,0,0,1,0,0,0,0,1,1,0,0,0)$. In this paper, the number of phonemes is 38. The equation (1) accumulates minimal distances along a speech interval after comparing the distances between a DPF vector $X_p(t)$ and DPF vectors of all the phonemes ($R_i$), and then select the $n$-th DPF vector that has minimal distance among extractors.

### 3.2. DPF extractor design based on clustering technique

In this section, a clustering procedure to divide spoken sentences into groups that are expected to represent some hidden factors is presented. The spoken sentence data which contains all the vowels of /a, i, u, e, o/ and either nasal-sound /m, n, N/ is used for clustering. The reason for including nasal sounds is that the individual difference of speech often appears in these sounds which are affected by many cavities of vocal organs. After clustering, a cluster to which each spoken sentence belongs is decided. The clustering procedure based on a Modified K-means algorithm [10] is done as follows:

1. Takes an average along each interval of all the same phonemes patterns. After averaging, at least six averages in /a, i, u, e, o, m, n, N/ are extracted from each spoken sentence data.
2. Selects two initial centroids that have maximum distance.
3. Selects two initial centroids that have maximum distance.
4. Repeats (3) until the number of clusters reaches to a preset value $N_c$. When dividing, the cluster with maximum variance is selected by normalizing accumulated distance of each cluster and comparing the normalized distances of all the clusters.

### 4. Experiments

#### 4.1. Experimental setup

##### 4.1.1. Speech database

The performance of the proposed method was evaluated using the Aurora-2J database [6]. The sampling rate was 8k Hz, and the utterances were connected Japanese digit strings. The database contains clean data as well as various types of noise-corrupted data. Different types of noise, for example, subway, babble, car, exhibition, restaurant, street, airport, station noise, are added/convolved in it. For the experiment in this paper, the training was performed using clean data only and the category was 0 (no change at back-end).

#### 4.2. Experimental setup

An input speech is sampled at 8 kHz and 25 ms Hamming-windowed speech segments are applied every 10 ms. For ES202 (W1008) performance, 12 Mel-cepstral parameters along with log power and their delta, and delta-delta parameters (total dimension 39) are used as feature vector.
11-vocabulary word-HMMs with 18 states are prepared together with a silent model with 5 states and a pause model with 3 states. In the proposed method, the number of clusters is varied from 1 to 8.

4.3. Experimental results

The experimental results of baseline (WI008) and the proposed DPF extractor with noise suppressor are shown in Table 1 and Table 2. Fig. 5 shows relative performances of WI008 and the proposed method comparing to WI007, which is a previous version of WI008 and does not include Wiener filter based noise reduction. In the experiment, the number of cluster N was fixed to four. From Table 1 and Table 2, we can see that the proposed method performed better than WI008 in clean and almost all noisy conditions. For example, in clean condition, the proposed method achieved 99.41% word accuracy, while WI008 had 98.48%. In noisy condition with SNR=0dB, the proposed method yielded word accuracy of 40.96%, while WI008 had 38.52%. Fig. 5 demonstrates the power of Wiener filter based noise suppression procedure. The proposed canonicalized DPF achieved 60.87% relative improvement over WI007, while WI008 gained it 59.09%. Between three data sets of A, B, and C, the C-set is open data concerning channel characteristics. This means WI008 is not effective on such distortion and thus we can know the proposed method gained most at the C-set (54.68%), while WI008 had 51.84% (see Fig. 5).

Table 3 shows the performance of the proposed method for different numbers of cluster. With four clusters, it showed the optimal performance. On the other hand, from the clustering point of view, we checked misclassification rate. The result of gender-type classification shows that only 10% data was misclassified. The clustering technique applied here is effective to the design of DPF extractors.

5. Conclusion

The design of multiple DPF extractors for canonicalization by using clustering technique was proposed. Because the proposed method isolates the feature extractor design from the HMM classifier design, it can be applied to DSR services. In the experiment on Japanese version AURORA2 database (AURORA2-J), the proposed method achieved significant improvements when combining the canonicalization process with the noise reduction technique based on a two-stage Wiener filter.

In future work, we will apply the canonicalization method to the other types of hidden factors and evaluate them in the practical environments.

Table 3 Overall average word accuracy [%] of the proposed method for different number of clusters.

<table>
<thead>
<tr>
<th>N (Number of clusters)</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Acc</td>
<td>75.92</td>
<td>77.71</td>
<td>78.94</td>
<td>78.04</td>
</tr>
</tbody>
</table>

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