Effect of frequency weighting on MLP-based speaker canonicalization

Yuichi Kubota¹, Motoi Omachi¹, Tetsuji Ogawa¹, Tetsunori Kobayashi¹, Tsuneo Nitta¹,²

¹Dept. of Compute Science, Waseda University, Tokyo, Japan
²Toyohashi University of Technology, Toyohashi, Japan

Abstract

Accurate and efficient speaker canonicalization is proposed to improve the performance of speaker-independent ASR systems. Vocal tract length normalization (VTLN) is often applied to speaker canonicalization in ASR; however, it requires parallel decoding of speech when estimating the optimal warping parameter. In addition, VTLN provides the same linear spectral transformation in an utterance, although optimal mapping functions differ among phonemes. In this study, we propose a novel speaker canonicalization using multilayer perceptron (MLP) that is trained with a data set of vowels to map an input spectrum to the output spectrum of a standard speaker or a canonical speaker. The proposed speaker canonicalization operates according to the integration of MLP-based mapping and identity mapping that depends on frequency bands and achieves accurate recognition without any tuning of mapping function during run-time. Results of experiments conducted with a continuous digit recognition task showed that the proposed method reduces the intra-class variability in both of the vowel and consonant parts and outperforms VTLN.

Index Terms: Speaker canonicalization, multilayer perceptron, feature extraction, connected digit recognition

1. Introduction

The performance of automatic speech recognition (ASR) is adversely affected by the inter-speaker variability, i.e., variations in time-frequency (TS) patterns of each phoneme. There are two efficient approaches to reduce this performance-degradation. One of the approaches is speaker adaptation that reduces acoustic mismatches [1, 2], and the other is speaker canonicalization in which input TS patterns are converted into speaker-invariant patterns.

Vocal tract length normalization (VTLN) [3, 4, 5] is often applied to speaker canonicalization. The frequency warping in VTLN can map an input onto a pattern that has few variation of the TS pattern. In general, VTLN needs parallel recognition stages with different warping parameters that are followed by maximum-likelihood optimization stages during run-time of the system i.e., the recognition step. This can cause undesirable delay in speech recognition. To avoid the processing time in parallel decoding, Emori et al. tried to directly estimate the warping parameter on the basis of a maximum-likelihood framework [6]. Most of these approaches, however, have applied the common linear mapping through a whole utterance period, even though the utterance includes different phonemes and the transfer function of a vocal tract is affected by both of speakers and phonemes. To handle this problem, VTLN has been modified so as to dynamically select the warping parameters frame-by-frame [7, 8]. This reduces recognition errors; however, the high computational cost of searching an optimum warping parameter is still required.

In this paper, we propose a novel speaker canonicalization method that uses MLP-based spectral mapping in which arbitrary speakers’ spectra are transformed into the canonical speaker’s spectra. Applying an MLP to speaker canonicalization derives two advantages over VTLN. First, once the spectral mapping function of MLP is trained, there is no need to optimize mapping function during run-time, that is, the computational cost during runtime can be reduced. Second, MLP can realize non-linear spectral mapping according to each phoneme in the utterance, and can achieve more accurate speaker canonicalization than VTLN. MLP-based spectral mapping has been studied in voice conversion, in fact, it applied to transform arbitrary speakers’ voices into a specific speaker’s voice [9, 10]. These works, however, have focused only on the perceptual assessment of synthesized speech, and have not yet discussed from the point of view of ASR performance. Since a vowel data set is only used to train the MLP-based mapping function on a frequency axis, the MLP outputs for consonantal parts are distorted in the low and high frequency bands, and the ASR performance was largely degraded. To handle these problems, frequency weighting and spectral integration to the outputs of the MLP is introduced and the weights at low, middle and high frequency bands are adjusted to improve the ASR performance. Experiments are conducted with a connected digit recognition task and compared with VTLN. The proposed speaker canonicalization method is expected to apply also for developing robust and efficient ASR systems by integrating with a speaker adaptive training framework [11, 12].

This paper is organized as follows. In Section 2, we describe the proposed speaker canonicalization method. Here, the effect of frequency weighting and spectral integration on MLP-based mapping is discussed. In Section 3, experiments conducted with a connected digit recognition task are explained. Finally, we present our conclusions in Section 4.

2. MLP-based speaker canonicalization

Figure 1 illustrates a schematic diagram of the proposed speaker canonicalization system. First of all, the centroid of speakers, or the canonical speaker, who is set for the target speaker in spectral mapping, has to be decided. In the present study, spectral mapping is carried out with a logarithmic mel-filter bank.
An MLP is introduced for the spectral mapping in which arbitrary speaker’s spectral patterns outputted from the filter bank is transformed to the canonical speaker’s one. During run-time, as described in Fig. 1, frequency weighting is applied to the spectra with and without MLP-based mapping and then the weighted spectra are integrated. Finally, mel-frequency cepstral coefficients (MFCCs) are extracted from the canonicalized spectra. In the rest of this section, we give more detailed explanation about canonical-speaker selection in 2.1, MLP-based spectral mapping in 2.2, and frequency weighting followed by spectral integration on the spectral mapping in 2.3.

2.1. Canonical speaker selection
The spectra of the canonical speaker are used for the target of spectral mapping in MLP training. The canonical speaker was chosen as a centroid of multiple speakers on the logarithmic mel-filter bank channel output space. In the present study, the canonical speaker was determined using only vowel data because the vowels mainly contribute to speaker characteristics [13], i.e., the spectra of vowels vary among speakers, while those of the consonants are not affected largely by the speakers. We used the JVPD corpus [14] to extract the canonical speaker. This corpus includes five Japanese vowels spoken by 385 persons (186 males and 199 females) whose ages range from 6 to 56 years old. First, one segment was chosen for each vowel and each speaker i.e., five segments per speaker. Then, 24–56 years old. First, one segment was chosen for each vowel.

2.2. MLP-based spectral mapping
Let $\tilde{x}(s,p)$ and $\tilde{y}(p)$ be the feature vectors for the phoneme $p$ spoken by the speaker $s$ and that spoken by the canonical speaker, respectively. Assume that a set of observed feature samples and their targets $\{\tilde{x}(s,p), \tilde{y}(p)\} \forall s,p$ for $p$ are given. A mapping function of $\tilde{x}(s,p)$ onto $\tilde{y}(p)$ for $\forall p$, which is described as $\tilde{y}(p) = f(\tilde{x}(s,p)) \forall s \forall p$, is trained with the three-layered MLP illustrated in Fig. 2. Here, $\tilde{x}(s,p)$ and $\tilde{y}(p)$ are obtained by concatenating three successive frames of 24-dimensional logarithmic mel-filter bank channel outputs, i.e., the center frame in the segment and its previous and subsequent frames. In this case, MLP-based speaker canonicalization can achieve higher representational spectral mapping than VTLN because a non-linear spectral mapping function is trained in consideration of the difference in phonemes, while in VTLN, common linear mapping is used, irrespective of phonemes comprising an utterance. There are 72, 144 and 72 units in the input, hidden, and output layers, respectively. In the present study, MLP was trained with only the vowel data set for the same reason as in selecting the canonical speaker.

During run-time, 24-dimensional logarithmic mel-filter bank channel outputs are extracted for each frame as $x_t \in \mathbb{R}^{24}$, irrespective of phonemes, and concatenated with the previous and subsequent frames’ outputs, consisting of a 72-dimensional vector as $\tilde{x}_t = [x_{t-1}, x_t, x_{t+1}]^T \in \mathbb{R}^{72}$. This vector is input to the MLP. In addition, this MLP-based mapping outputs a 72-dimensional vector for each frame as $\tilde{y}_t \in \mathbb{R}^{72}$. 24-dimensional components in the middle of $\tilde{y}_t$ are chosen as $y_t \in \mathbb{R}^{24}$ and then analyzed with DCT to extract MFCCs.

In this case, we have several options for the MLP structure. In fact, we investigated the case of 72–144–24 units and that of 72–144–72 units. The preliminary experiments showed that the latter system gave better accuracy, and, therefore, it was used in the present study.

2.3. Frequency weighting and spectral integration
The aim of MLP-based spectral mapping using vowel data is to extract the frequency-warping function that can map input spectrum onto the canonicalized spectrum; however, it can induce undesirable effects on consonantal parts. Figure 3 shows spectra of vowel and consonantal parts with and without MLP-based spectral mapping. Here, leave-one-out MLP training and testing conducted using 385 samples for each vowel $v$, $\{(x(v,v), y(v,v))\}_{v=1}^{385}$, from the JVPD corpus, i.e., 384 for MLP training and one for testing. Figures 3(a) and 3(b) show that as for the vowel, spectral mapping of arbitrary speakers onto the canonical speaker can effectively reduce the intra-class variability induced by the difference in speakers. The fundamental frequency of vowels spoken by the canonical (female) speaker, however, exists in the low frequency band, e.g., around 300 Hz, and a spectral peak appears at that frequency. In this case, the spectral mapping function tends to be trained such that the formant frequency would be unduly fit to the spectral peak derived from the fundamental frequency of the canonical speaker. In addition, from Figs. 3(c) and 3(d), the spectra of consonantal parts are degraded through MLP-based spectral mapping, especially in the low and high frequency bands, because MLP-based spectral mapping is trained with only vowels.

To improve these problems, we attempt to apply frequency weighting and spectral integration to the outputs of MLP. The outputs from MLP-based mapping can be used as the canonicalized spectra in the middle frequency band and, at the same time, the negative impact of MLP-based mapping should be reduced in the low and high frequency bands. The frequency weighting...
Figure 3: Spectra with and without MLP-based spectral mapping. Bold line represents spectrum from canonical speaker. (a) and (c) show spectra of vowel /a/ and consonant /h/ without MLP-based mapping, respectively. (b) and (d) show spectra of /a/ and /h/ with MLP-based mapping, respectively.

Figure 4: Frequency weighting functions for inputs and outputs of MLP-based spectral mapping. $w^{\text{out}}(k)$ and $w^{\text{in}}(k)$ express weighting functions for spectra with and without MLP-based mapping, respectively. Here, gradients of those functions at $k_L$ and $k_H$ are $\pm(1-\alpha)/4$.

Figure 5: Spectra with and without speaker canonicalization applying frequency weighting and spectral integration to MLP-based mapping. Bold line represents spectrum from canonical speaker. (a) and (c) show spectra of vowel /a/ and consonant /h/ without spectral mapping, respectively. (b) and (d) show spectra of /a/ and /h/ with spectral mapping, respectively.

3. Connected digit recognition experiment

Experimental comparisons were made for connected digit recognition. We compared the digit accuracy of three features.

1. **MFCC**: MFCCs without speaker canonicalization
2. **VTLN**: MFCCs with VTLN
3. **MLP-FWSI**: MFCCs with MLP-based spectral mapping to which frequency weighting and spectral integration are applied.
3.1. Experimental setup

3.1.1. Speaker canonicalization

Experimental conditions for speech analysis are listed in Table 1. In MLP training, 24-channels of logarithmic mel-scale filter bank outputs are calculated for vowels from arbitrary speakers and those from the canonical speaker and then concatenated with the previous and subsequent frames’ outputs, consisting of 72-dimensional vectors. The MLP was trained with 1,925 vowels from the JVPD corpus in which 5 vowels are spoken by each of 385 speakers. One of the 385 speakers was assigned to be the canonical speaker. The activation function used in the MLP is a logistic sigmoid function with a gradient coefficient of 0.3. The learning rate was fixed to 0.7. The MLP weights were updated in the online learning manner, i.e., for each example. \( \alpha \) and \( k_H \) described in Fig. 4 were 5 and 19, which correspond to 400 Hz and 4000 Hz, respectively. They were arbitrarily determined from the preliminary experiment. Digit recognition accuracy was evaluated with \( \alpha = 0.0, 0.1, 0.2, 0.3, \cdots, 1.0 \).

3.1.2. Speech recognition

The speech data were sampled at 16 kHz and quantized into 16 bits of data. It should be noted that VTLN and the proposed speaker canonicalization were applied to the data for training acoustic models and those for testing. The acoustic feature parameters were 26-dimensional and consisted of 12-dimensional MFCCs, a power, 12-dimensional ΔMFCCs, and a Δpower. The cepstral mean normalization was computed and applied to each utterance. The frame length and frame shift were 30 and 10 ms, respectively.

We used a gender-independent monophone HMM. The distribution function in each state of the models was represented by a 16-mixture Gaussian distribution with diagonal covariances. These models were trained with 41,396 sentences taken from the ASJ database [15], which includes sentences from Japanese newspaper articles and phoneme-balanced sentences. The evaluation data comprised 4,004 utterances of connected digits spoken by 104 speakers. The connected digits comprised eleven Japanese digits (1-9 plus two words for “0”): “ichi,” “ni,” “san,” “yon,” “go,” “roku,” “nana,” “hachi,” “kyu,” “zero,” and “maru,” as adopted by CENSREC-1 [16].

3.2. Experimental result

Figure 7 describes the digit accuracy obtained from the connected digit recognition system employing the proposed MLP-based speaker canonicalization as a function of \( \alpha \) defined in Fig. 4, which expresses the value of \( w^{out}(k) \) in the low and high frequency channels. Those results are compared with the digit accuracy obtained from the system using MFCC without any speaker canonicalization and that using MFCC with VTLN. \( \alpha = 0.0 \) means that MLP-based spectral mapping is not used at all in the low and high frequency channels. \( \alpha = 1.0 \) means that MLP-based mapping is used for all frequency channels. MLP-based spectral mapping without frequency weighting and spectral integration (\( \alpha = 1.0 \)) increased the number of digit errors of the system without any speaker canonicalization (MFCC). In contrast, applying frequency weighting and spectral integration to MLP-based mapping (MLP-FWSI) achieved the best accuracy in the case of \( \alpha = 0.1 \), reducing the digit error rate by 3.6 % compared with MFCC and 1.6 % compared with VTLN. These results showed that the proposed method achieved accurate speaker canonicalization without training of consonant data by using frequency weighting and spectral integration.

4. Conclusion

We proposed a MLP-based speaker canonicalization method with frequency weighting and spectral integration to reduce the intra-class variability. The proposed method could achieve accurate speaker canonicalization while avoiding undesirable effects induced by MLP-based mapping in the low and high frequency bands. From connected digit recognition experiments, the proposed method improved digit accuracy by 3.6 % and 1.6 % compared with the case without speaker canonicalization and VTLN, respectively. Investigating the effect of proposed canonicalization on LVCSR is one of near future works.

5. Acknowledgements

This work has been supported by Strategic Information and Communications R & D Promotion Programme by MIC, Japan.
6. References


