Automatic speech pronunciation correction with dynamic frequency warping-based spectral conversion

Nobukatsu Hojo, Hirokazu Kameoka, Kou Tanaka, Takuhiro Kaneko

NTT Communication Science Laboratories, NTT Corporation, Japan
hojo.nobukatsu@lab.ntt.co.jp

Abstract—This paper deals with the problem of pronunciation conversion (PC) task, a problem to reduce non-native accents in speech while preserving the original speaker identity. Although PC can be regarded as a special class of voice conversion (VC), a straightforward application of conventional VC methods to a PC task would not be successful since with VC the original speaker identity of input speech may also change. This problem is due to the fact that two functions, namely an accent conversion function and a speaker similarity conversion function, are entangled in an acoustic feature mapping function. This paper proposes dynamic frequency warping (DFW)-based spectral conversion to solve this problem. The proposed DFW-based PC converts the pronunciation of input speech by relocating the formants to the corresponding positions in which native speakers tend to locate their formants. We expect the speaker identity is preserved because other factors such as formant powers are kept unchanged. In a low frequency domain evaluation results confirmed that DFW-based PC with spectral residual modeling showed higher speaker similarity to original speaker while showing a comparable effect of reducing foreign accents to a conventional GMM-based VC method.

Index Terms—Accent conversion, dynamic frequency warping, voice conversion

I. INTRODUCTION

This paper deals with the problem of pronunciation conversion (PC), a problem to reduce accent in speech while preserving the speaker identity of the original speech. Accents are differences in pronunciation by a community of people from a national or regional geographical area, or a social grouping [1]. It is known that differences by accents are manifested in the differences such as the formants and their trajectories [2], [3] or pitch intonation and duration parameters [4], [1]. Reducing these accents from speech while preserving the original speaker identity will be beneficial for applications such as language educational systems for second language learners [5], [6], [7] and teleconference scenarios where people with different nationalities participate.

If we consider an accent as one of non-para-linguistic information, PC can be regarded as a special class of a voice conversion (VC) problem, a problem to convert non-para-linguistic information while preserving linguistic information. For VC, data-driven statistical methods have been successfully introduced during the last two decades [8], [9], [10]. One successful VC method is based on Gaussian mixture models (GMMs) [9], [10]. Most of the conventional VC methods first train a mapping function between the acoustic features of source and target speech using parallel data, i.e. a pair of time-aligned feature sequences of source and target speech. At test time, a feature sequence of the input speech is converted using the trained mapping function. A direct application of these methods to the PC task would not be successful because the original speaker identity may also change. This problem is because two functions, the accent conversion function and the speaker identity conversion function, are entangled in an acoustic feature mapping function.

This work is based on a belief that a dynamic frequency warping (DFW)-based spectral conversion approach can be a reasonable solution to this problem. The DFW-based spectral conversion approach was originally proposed [11], [12] mainly for the purpose of improving naturalness of converted speech. Since one of the dominant differences in accented speech appears in formant frequency trajectories [2], [3], we expect that pronunciation can be corrected via frequency warping if the formants can be relocated to the positions in which native speakers tend to locate their formants. In this way, we expect that the original speaker identity will not be affected since DFW does not convert the other factors such as formant powers and spectral tilts.

The main purpose of this work is to investigate the effectiveness of DFW-based spectral conversion for the PC task. Furthermore, there are two points that we want to investigate. One is the effectiveness of spectral power interpolation. DFW only allows us to deform source spectra in the frequency direction and does not have an ability to reduce the gap in spectral powers. Since the peakiness of each formant is also expected to be an important factor that characterizes pronunciation quality, we would also want to deform source spectra in the power direction. However, if we convert source spectra exactly to target spectra, the speaker identity will no longer be preserved. Thus, there is a trade-off between the pronunciation quality and the voice similarity to a source speaker. This trade-off is investigated by subjective evaluation. The other point is concerned with the modeling of frequency warping functions. The idea of the method in [12] is to associate a frequency warping function with each Gaussian of the GMM that models a joint distribution of source and target spectra. The frequency warping function is obtained from a pair of source and target spectra averaged over all the frames assigned to the same Gaussian. We found that the averaged spectra tend to be over-smoothed and so the obtained warping function will also be over-smoothed. To avoid this, we propose a method that first extracts frequency warping functions from pairs of source and target spectra frame-by-frame, treats the obtained frequency warping functions as features to be predicted, and models the joint distribution of source spectra and the frequency warping
functions using a GMM. In this paper, we particularly focus on the problem of spectral conversion only. However, it can be used in combination with prosody conversion methods [5], [13].

II. DYNAMIC FREQUENCY WARPING-BASED ACCENT CONVERSION

The proposed method consists of training process and conversion process. We show the overall architecture of the proposed method in Fig. 1.

A. Dynamic frequency warping with frequency derivative distance

The proposed method first finds an optimal warping of frequency axis with DFW [14], [15]. Let us denote the time aligned source spectra by \( X = \{x_t\}_{t=1}^F \) and the target spectra by \( Y = \{y_t\}_{t=1}^F \). Here \( x_t = [x_{f,t}]_{f=1}^P \) and \( y_t = [y_{f,t}]_{f=1}^P \) respectively denote the source and target spectrum, \( t, f \) denote the frame and frequency indices, respectively. The warping function \( \hat{w}_t = [w_{f,t}]_{f=1}^P \), which we call the DFW vector, can be obtained as the path that minimizes the spectral distance between a frequency warped source spectrum and a target spectrum;

\[
\hat{w}_t = \arg \min_{w_1, \ldots, w_F} \sum_{f=1}^F D(x_{w_{f,t}}, y_{f,t}),
\]

where \( w_f \) takes a frequency index \( w_f \in \{1, \ldots, F\} \) in an ascending order. By restricting the warping path to be \( w_{f+1} = w_f \in \{0, 1, 2\} \) for each \( f \in \{1, \ldots, F-1\} \), a smoothed and fixed length DFW vector can be extracted. Although it is possible to use the \( l_2 \) norm of a log spectral difference to define \( D(x, y) \), the obtained warped spectra tend to have plateau because of the power difference of spectral peaks between source and target speakers. We concern that the plateau can degrade harmonics and quality of converted speech, which is the reason we define the distance introducing the frequency derivative distance term as follows in this work;

\[
D(x_f, y_f) = ||\log x_f - \log y_f||_2 + \gamma ||\dot{x}_f - \dot{y}_f||_2,
\]

\[
\dot{x}_f = \log x_{f+1} - \log x_f,
\]

\[
\dot{y}_f = \log y_{f+1} - \log y_f
\]

where and \( \gamma \) indicates the weight for the frequency derivative term. This simple spectral distance is expected to have a similar effect to correlation based DFW [16] or DFW based on the histogram of spectral peaks [17]. An example of spectral warping using frequency derivative distance is illustrated in Fig.2. We can see that the spectral plateau is eliminated by introducing the frequency derivative distance term.

Since we would want to eliminate the speaker identity information from the DFW vector as much as possible, we found it necessary to apply vocal tract length normalization (VTLN) [18] to the source and target speech spectra before performing DFW.

B. GMM training and DFW vector estimation

For estimating DFW vectors from the source spectra, we model the joint distribution of the source spectra and the DFW vectors using a GMM. In order to avoid overfitting in modeling high dimensional vectors, we compress the dimension of the feature vectors and constitute a joint vector \( z_t = [m^T_t, d^T_t]^{T} \), where \( m_t \) is the mel-cepstrum extracted from the source spectrum \( x_t \) at frame \( t \). The vector \( d_t \) is derived as follows;

\[
d_t = \text{DCT}(w_t - w_b),
\]

where \( w_b = [1, \ldots, F]^T \) and \( \text{DCT}(\cdot) \) denotes the discrete cosine transform. We model the joint feature as follows;

\[
p(z) = \sum_{i=1}^I \alpha_i N(z; \mu_i, \Sigma_i), \sum_{i=1}^I \alpha_i = 1, \alpha_i > 0,
\]

where \( N(z; \mu, \Sigma) \) denotes the normal distribution with the mean vector \( \mu \) and the covariance matrix \( \Sigma, \alpha_i \) denotes a weight of class \( i \), and \( I \) denotes the total number of the Gaussian mixtures. Given a training set of \( \{m_t, d_t\}_{t=1}^T \) pairs, we train the GMM parameters using the EM algorithm.
At test time, DFW vectors are estimated using the trained GMM and the input spectra. The mapping function [10] is given by

$$F(m) = E[d|m]$$

$$= \sum_{i=1}^{I} h_i(m) \mu_i^{(d)} + \Sigma_i^{(dm)} (\Sigma_i^{(mm)})^{-1} (m - \mu_i^{(m)})]$$

(7)

$$h_i(m) = \frac{\alpha_i N (m; \mu_i^{(m)}, \Sigma_i^{(mm)})}{\sum_j \alpha_j N (m; \mu_j^{(m)}, \Sigma_j^{(mm)})},$$

(8)

where $\mu_i^{(m)}$ and $\mu_i^{(d)}$ denote the mean vectors of class $i$ for the mel-cepstra and DCT of DFW vectors. The estimated DFW vector is derived as follows;

$$\hat{w}_t = iDCT(F(m_t)) + w_b,$$

(9)

where $iDCT(\cdot)$ denotes the inverse discrete cosine transform. Since the estimated $\hat{w}_t$ takes a continuous value, they are floored to be an integer frequency index to convert spectra as $\hat{y}_t = [x_{\hat{w}_t,t}]F_{t=1}^T$.

C. Spectral residual modeling

Note that DFW only has an ability to deform source spectra in the frequency direction and does not have an ability to fill the gap in spectral powers. Since the power of each formant is also expected to be an important factor that characterizes pronunciation quality, we would also want to deform source spectra in the power direction. However, if we convert source spectra exactly to target spectra, the speaker identity will no longer be preserved. Thus, there is a trade-off between the pronunciation quality and the voice similarity to a source speaker. Here, we also consider predicting the power differences between the warped spectra and the target spectra so that we can add the predicted differences to the warped spectra at test time.

The residual spectra $r_{f,t} = [r_{f,t,t}]F_{t=1}^T$ is defined as the difference between the target spectra and the warped source spectra as follows;

$$r_{f,t} = \frac{y_{f,t}}{x_{\hat{w}_t,t,t}}$$

(10)

This residual spectra is extracted for each frame and construct a joint vector $s_t = [m_t, q_t]$, where $q_t$ denotes the DCT of $\log r_t$. This joint vector is modeled by another GMM and used to estimate residual spectra $\hat{r}$ in the same manner for estimating the residual vector. The output converted spectra $\tilde{y}_{t}^{(r)} = [\tilde{y}_{f,t,t}]F_{t=1}^T$ is derived as follows;

$$\tilde{y}_{f,t}^{(r)} = \tilde{y}_{f,t} \cdot \tilde{r}_{f,t}$$

(11)

where $\lambda$ denotes the weight for spectral residual modeling.

III. EXPERIMENTAL EVALUATION

A. Experimental conditions

We evaluated pronunciation similarity and speaker identity of the converted speech to compare the performance of the conventional VC and proposed methods. We used an Indian male speaker as the source (hereafter, the SRC) and an American male speaker as the target (the TGT). The dataset consisted of 65 parallel utterances (5.2 minutes). We used 20 utterances for training and the others for training. Table I shows the statistical information of the corpus. The data were sampled at 16kHz, then 25 mel-cepstral coefficients, fundamental frequency ($F_0$), and aperiodicities were extracted every 5ms by using the STRAIGHT analysis system [19]. To obtain parallel utterances, we used dynamic time warping (DTW) to align mel-cepstral sequences of the source and target speakers [20]. We evaluated and compared the following 6 spectral conversion methods.

- GMM-MCEP: The conventional GMM-based VC method using mel-cepstral feature $s$ [10].
- DFW-CENTROID: DFW-based spectral conversion method. DFW functions were derived from centroid spectra pair of each Gaussian.
- PROP-0: The proposed method without spectral residual modeling ($\lambda = 0.00$).
- PROP-1: The proposed method with spectral residual modeling ($\lambda = 0.33$).
- PROP-2: The proposed method with spectral residual modeling ($\lambda = 0.67$).
PROP-3: The proposed method with spectral residual modeling ($\lambda = 1.00$).

Although DFW-CENTROID was implemented emulating the conventional DFW-based VC method [12], details such as vocoder for synthesizing speech was not the same. This is in order to reduce the factors considered in the experiments. We evaluated 4 proposed methods with different spectral residual weight in order to evaluate the trade-off of the pronunciation quality improvement and speaker identity degradation.

For DFW-CENTROID, we first train a GMM that models joint distribution of source and target spectra using line spectral pairs (LSP) features. Then, a pair of source and target centroid spectra was reconstructed from the mean vector of each Gaussian. The frequency warping function was obtained from the pair of source and target centroid spectra using the DFW method in sec. II-A. The conversion procedure was similar to the proposed method except for using LSPs instead of mel-ceptra as input spectral features. We used 25 dimensional LSP for this method.

For the proposed methods, we used 25 dimensional mel-ceptra and 25 dimensional DCT of DFW vectors. We used mel-ceptra for DFW vector extraction so that the warping function gets higher frequency resolution in a low frequency domain. We set $\gamma = 25.0$ in eq. (2), which was determined experimentally. We show the results without conducting VTLN before DFW vector extraction, because the results were not improved. We consider this is because the vocal tract length before DFW vector extraction, because the results were not experimentally. We show the results without conducting VTLN domain. We set $\gamma = 25.0$ in eq. (2), which was determined experimentally. We show the results without conducting VTLN before DFW vector extraction, because the results were not improved. We consider this is because the vocal tract length before DFW vector extraction, because the results were not improved.

For all of the methods, the number of mixture components for GMMs was 16. All of the speech samples were synthesized by STRAIGHT vocoder [19], using converted spectra by each method, F0 and aperiodicity of the SRC speaker.

Fig. 3 shows the example of warping functions reconstructed from $\mu^{(d)}$ of the trained GMM of PROP-0. We can see that these functions can warp frequency by 200 Hz around 1–3 kHz frequency region, which we expect is enough warping width to relocate formants to reduce accents.

**IV. CONCLUSION**

This paper investigated the effectiveness of DFW-based spectral conversion for pronunciation conversion task, a prob-
lem to reduce accent in speech while preserving the speaker identity of the original speech. The proposed method converts the pronunciation of input speech by relocating the formants to the corresponding positions in which native speakers tend to locate their formants. We expect the speaker identity is preserved because other factors such as formant powers are kept unchanged. Subjective evaluation results confirmed that DFW-based pronunciation conversion with spectral residual modeling showed higher speaker similarity to original speaker compared to a conventional GMM-based VC method. It is worth investigating whether utilizing neural network-based methods [21], [22], [23] for DFW vector estimation is effective with larger size of speech corpus. Future works include extending the experiments to other speaker, accent and language pairs.

REFERENCES