ABSTRACT

A parametric conversion of speech individuality is proposed based on STRAIGHT speech representation. STRAIGHT speech analysis-synthesis can produce high quality speech for various kinds of transformations by using 1) pitch synchronous windowing, 2) time-frequency spectrum interpolating and 3) randomized all-pass filtering for shaping phase spectrum. In order to utilize the smoothness of STRAIGHT spectrum, speech conversion is accomplished by warping the frequency axis. The warping functions are trained for each class of the predetermined spectrum shape grouping. The evaluation test is performed to compare the proposed method and VQ prototype mapping or linear transformation of cepstrum vectors. As a measure of converted speech quality, the MOS score of 6 subjects is calculated and is found to be better than conventional methods by about 1.5 point without degrading the accuracy of speech individuality discrimination.

Keywords: Speaker Individuality, STRAIGHT, warping frequency axis.

1. INTRODUCTION

Controlling the speech individuality is one of the most important issues in extending the application fields of natural dialogue systems. Especially, in amusement and education applications, generating multi-speaker's speech in natural quality is strongly required. As Kuwabara and Sagisaka classified in [1], the parametric controlling and the parameter mapping have been the two major approaches of speaker conversion. Since it is not easy to extract and control acoustic parameters related to speech individuality, to find a mapping function between physical parameters such as spectrum and cepstrum, becomes more common recently. The one of the most successful mapping approaches was fully non-parametric one, codebook mapping proposed by Abe et al.[2]. To improve naturalness as well as speaker individuality of synthetic speech, a parametric mapping such as vector interpolation, linear transformation [3] and frequency scaling [4] are also studied. However, less attention is paid to the speech representations on which the mapping is defined.

In this paper, we discuss on the application of new speech representation STRAIGHT [5] to speaker conversion. Since STRAIT is basically a vocoder, its flexibility on signal manipulation is suits not only for finding appropriate mapping, but also for synthesizing speech.

2. STRAIGHT SPEECH ANALYSIS-SYNTHESIS SYSTEM

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectrogram) is a newly proposed speech analysis-synthesis system which is characterized by smoothed spectrum representation and fine flexible controlling of excitation. It is reported that, in STRAIGHT framework, the naturalness of the speech quality can be kept in up to 600% change of parameters, e.g. pitch frequency or speech rate. The basic scheme of the STRAIGHT analysis-synthesis system used in this paper can be summarized as follows.

1. Calculate smoothed spectral pattern $S(\omega,t)$ by applying interpolating function on temporal spectral pattern $F(\omega,t)$. For the interpolating function, a two dimensional gaussian function is used.
2. Complex cepstrum representation is calculated from the smoothed spectrum.
3. Applying a set of all-path filters to the reconstructed minimum-phase impulse response waveform. Each all-path filter is carefully controlled so that the pitch period is strictly preserved as well as the phase of the reproduced sound has a natural randomness.

3. NONLINEAR FREQUENCY WARPING

3.1 SPECTRUM MAPPING THROUGH NONLINEAR FREQUENCY WARPING

The basic idea of using frequency warping for speaker conversion is same as [4]. That is to find an optimal frequency warping functions for classes of speech spectrum shapes by

\[
\theta_{s}(f) = \arg \min_{\theta} \int_{0}^{\infty} (\tilde{S}_{x}(f) - \tilde{S}_{y}(\theta(f))) df ,
\]

where \(\tilde{S}_{x}(f)\) and \(\tilde{S}_{y}(f)\) are the average power spectrum functions of original and target speakers averaged over training samples falling into the same speech spectrum shape class. The power spectrum functions are calculated from smoothed STRAIGHT spectrum. Unlike [4], in the higher frequency region, the tilt of the mapping function is fixed. In other words, the endpoint of the integration in (1) is limited to the lower frequency region, so as to disregard the differences of the spectrum in the low energy region. The tilt, \(\alpha\), is determined by the ratio of pitch frequencies of the original and target speakers, i.e.,

\[
\alpha = \tan^{-1} \left( \frac{f_{y} - f_{x}}{n_{y} - n_{x}} \right) ,
\]

which was found to be best in a speech conversion within a speaker. For the same reason, the global tilt removal, which emphasizes the difference in higher frequency region was not performed. However, the difference spectrum between the VQ prototypes of the original and its corresponding target speakers are stored as a residual spectrum.

3.2 SYNTHESIS

The synthesis procedure is illustrated in Figure 1. In the synthesis step, 1) prototype vectors of original speaker for classification, 2) residual spectrum vectors for all classes and 3) frequency warping functions for all classes, are required. The outline of the procedure can be summarized as follows.

1. Calculate smoothed spectrum and its cepstrum of input (original speaker's) speech by means of STRAIGHT analysis.

2. Classify input speech, in frame-by-frame basis, into one of predetermined classes based on cepstral distance measure.

3. By applying the frequency mapping function defined for the class, convert the original smoothed spectrum of the frame to the new spectrum. Before applying, the warping function is smoothed by a triangular time window,

\[
\theta(f;n) = \frac{\sum_{i=-N}^{N} (N-|n-i|) \theta(f;n-i)}{\sum_{i=-N}^{N} |n-i|} ,
\]

where \(n\) is the time index of the processing frame. The residual spectrum is also smoothed and added to the converted spectrum.

4. Linearly transform pitch frequency in log domain

\[
\frac{\ln f_{y} - \mu_{x}}{\sigma_{x}} = \frac{\ln f_{x} - \mu_{y}}{\sigma_{y}} ,
\]

where \(\mu\) and \(\sigma\) are mean and standard deviation of the logarithm of pitch frequency.

5. Finally, reconstruct the target speech by STRAIGHT synthesizer.

3.3 TRAINING

The training algorithm of mapping functions and residual spectrums is illustrated in Figure 2. The algorithm can be summarized as follows.

1. Calculate STRAIGHT smoothed spectrum, then find frame-by-frame correspondence between original and target speech of training utterances by DTW.

2. Classify all frames of original speech into one of the predetermined classes. VQ codebook for the classification is designed before training. The target speech for training is also classified into the groups based on the class of the corresponding original speaker's frame.

3. Calculate the average spectrums of the target speech classes.

4. Find an optimal warping function of frequency by matching the average spectrums of the original and the target of each class. Store the found optimal warping function for each class of origi-
nal speech spectrum. The residual spectrum, difference between the target and the modified original spectrums, after the optimal warping is also stored for each class.

4. EVALUATION TEST
The proposed method is evaluated from the viewpoints of the speech quality and the speech individuality through comparing with codebook mapping [2] and linear transformation of cepstrum vectors [3]. The numbers of mapping functions, i.e., number of codewords in codebook mapping and number of transformation matrices in linear transformation, are 512 and 128, respectively. Those numbers are decided by preliminary experiments. For the proposed method, the number of warping functions is changed from one to 1,024. Two male and two female speakers are used for the test. Each speaker uttered fifty phonetically balanced sentences; forty-eight sentences are used for training and rest two sentences are used for the tests. All utterances are digitized into 16 bits by 16 kHz sampling frequency. Frame shift is fixed to 5 ms, whereas the frame length is determined by the pitch period (number of point for FFT analysis is fixed to be 1024). For the clustering of mapping functions, 40 cepstrum parameters derived from smoothed spectrum of STRAIGHT are used.

4.2 SPEECH QUALITY
The subjective quality of the converted speech of each method is quantified by MOS value of five male subjects. The results are shown in Figure 3. Comparing with codebook mapping and linear transformation, the subjective quality of frequency mapping is better by about 1.5 point. Furthermore, degradation due to increasing the mapping class is not so severe as of the linear transformation of cepstrum vectors.

4.2 SPEECH INDIVIDUALITY
The speaker individuality of the converted speech is also evaluated by the speaker discrimination experiments. In the experiment, subjects answer that converted speech sounds more like original or target speaker. The ratio of the answer that the converted speech sounds more like target speaker is shown in Figure 4. In female to female conversion, the frequency warping of STRAIGHT spectrum outperforms codebook mapping. The correct rate of the converted speech by linear transformation of cepstrum vectors, which is not given here, is slightly lower than codebook mapping. In male to male conversion, however, there is no significant difference among them.

5. CONCLUSION
A speaker conversion method based on nonlinear frequency warping of STRAIGHT speech representation is discussed. In the speech quality evaluation, when compared with codebook mapping and linear transformation of cepstrum vectors, more than 1.5 point better MOS value is obtained. In the test, the better results are obtained when the more mapping functions are utilized. In the speech individuality discrimination test, the proposed method also outperformed the codebook mapping and linear transformation. However, both speech quality and individuality are still insufficient. It was found that to find appropriate frequency warping is sometime very difficult. Improving the spectral measure for finding optimal warping is the most important future work.

ACKNOWLEDGEMENT
This work has been partially supported by CREST program of JST, Japan.

REFERENCES
Figure 1: Speech conversion procedure.

Figure 2: Training procedure.

Figure 3: Subjective quality scores of the speaker converted speech. CM stands for Codebook Mapping, LC for Linear transformation of Cempstrum vectors.

Figure 4: Correct rate of speech individuality identification test. CM stands for Codebook Mapping.