CONTINUOUS TIME-FREQUENCY COORDINATE MAPPING
WITH SPARSE ANCHORING TEMPLATES AND
ITS APPLICATION TO AUDITORY MORPHING

PACS: 43.55.Cs

Takahashi, Toru; Irino, Toshio; Kawahara Hideki
Faculty of Systems Engineering, Wakayama University, 930 Sakae-dani, Wakayama, Japan;
tall@sys.wakayama-u.ac.jp, irino@sys.wakayama-u.ac.jp, kawahara@sys.wakayama-u.ac.jp

ABSTRACT

The concept of unified speech style manipulations based on STRAIGHT is referred to as voice texture mapping. This powerful tool has potential for speech research and the post production of speech materials. Voice morphing is one implementation of voice texture mapping. Highly natural morphing, which has been widely used in various applications involves transforming reference and target parameters represented on each time-frequency axis into a common time-frequency axis. However, the transformation function was modeled on a segmental bilinear function, and segment boarders were manually assigned in the original STRAIGHT-based morphing. This manual pre-processing made the original STRAIGHT-based morphing time-consuming.

This paper proposes a new method for designing time-frequency coordinate transformation functions without relying on reference point assignments. Transformation functions in the proposed method are generated from automatically generated spectrum mapping and level compensation functions.

Speech sounds morphed by the proposed method are compared with speech samples morphed by the original STRAIGHT-based morphing. To evaluate the proposed method, ABX tests were conducted, and test results showed that the proposed method was limitedly effective to morph speech samples. The morphed speech samples are available at (http://www.sys.wakayama-u.ac.jp/~tall/indexe.html).

INTRODUCTION

The concept of unified speech style manipulations based on STRAIGHT [1] is referred to as voice texture mapping [2]. The powerful tool has potential for speech research and the post production of speech materials. Auditory morphing is one implementation of voice texture mapping. Speech morphing is auditory morphing specialized in speech processing. Highly natural speech morphing [3] has been widely used in various applications [4, 5]. Speech morphing is a procedure that generates intermediate synthetic speech from two exemplar speech samples. It is necessary to provide a means to align the time-frequency coordinates of the reference and target exemplar speech samples. If alignment is omitted, the morphed speech is severely distorted. Once the reference and target parameters are represented on a common time-frequency axis, the parameters are linearly combined. Finally, morphed speech sound is re-synthesized from the combined parameters.

Speech morphing can be considered a special case of a voice texture mapping. The voice texture parameters are composed of reference and target speech parameters in speech morphing. The texture parameters are mapped with some weight into the reference speech parameters. The weight corresponds to the morphing ratio. The voice texture parameters are represented as the difference between the time-frequency parameters, which are transformed to compensate for the differences of spectral peak position. The parameters need time-frequency transformation functions to make the voice texture parameters. However, the morphing procedure consists of the manual assignment of reference points for designing time-frequency coordinate transformation functions. Such manual pre-processing made STRAIGHT-based morphing time-consuming and very difficult for general users.
This paper describes an automatic speech morphing method and proposes an automatic time-frequency coordinate transformation function designing method without relying on reference point assignments. The designing method also enables automatic mapping of voice texture. The method consists of bilateral voice style conversion, whose transformation functions are generated from automatically generated spectrum mapping and level compensation functions.

In the next section, STRAIGHT is briefly described. Then speech morphing based on bilateral voice style conversion is proposed in Section 3. Experimental results are shown in Section 4. Finally, this paper is concluded.

STRAIGHT

STRAIGHT is a variant of a channel Vocoder that analyzes input speech signals into three parameters: fundamental frequency, spectrogram (STRAIGHT spectrogram), and aperiodicity index. STRAIGHT parameters are represented in terms of non-negative numbers. The STRAIGHT spectrogram is characterized by virtually complete elimination of interferences due to periodicity from spectrographic representation using a complementary set of time windows. STRAIGHT parameters, which are smoothed spectrum but preserve spectral peak information enable flexible manipulations of each parameter. The aperiodicity index, which represents the amount of random components in a given time frequency-region, is used to control the mixed mode (pulse, pulse noise) excitation source.

SPEECH MORPHING BASED ON BILATERAL VOICE STYLE CONVERSION

Issues in STRAIGHT-based morphing using anchoring points

The original STRAIGHT-based morphing relies on the manual assignment of representative time-frequency locations (anchoring points) to be aligned to define the time-frequency coordinate transformation by anchoring points because the time-frequency coordinate transformation is modeled by segmental bilinear functions. Once this transformation is defined, STRAIGHT-based morphing is implemented using the linear interpolation of each parameter represented on the reference coordinates followed by time-frequency coordinate mapping to the morphed coordinate. This manual assignment of anchoring points is the source of difficulties.

The first issue is its excessive flexibility. The underlying reason for using anchoring points to define the necessary coordinate transformation was originally intended to provide researchers complete control of transformations. This freedom would be desirable for psychoacoustical experiments that require precise control of physical parameters but it is excessively flexible for general speech applications.

The next issue is the ambiguity of anchoring points. A coordinate transformation function must be designed to align two time-frequency representations. Anchoring points are not uniquely determined to define a set of effectively relevant functions.

The proposed method alleviates these issues by providing means to design to directly design necessary transformation functions primarily using vowel-based information.

Outline of voice style conversion based on vowel information

An overview of the proposed speech morphing based on bilateral voice style conversion is displayed in Fig. 1. The proposed method can be implemented as an automatic process, since automatic voice style conversion has already proposed [6]. In Fig. 1, the morphing source and target are indicated as A and B. These are simultaneously the reference speeches of voice style conversions. The voice styles of A and B can be independently converted into another style to morph reference speeches A and B. Conversion degree $\alpha$ is introduced into the voice style conversion. When morphing source A and morphing target B are morphed at a ratio of 0.4, A is converted into B at weight 0.4 and B is converted into A at weight 0.6 (given by 1-0.4). After that, morphed speech parameters are given by morphing those converted parameters at a ratio of 0.4.

Conversion degree $\alpha$ is introduced into the voice style conversion. STRAIGHT spectrogram conversion consists of two steps: spectral warping and level compensation. Aperiodicity indices are morphed in a similar manner. Before applying to be morphed, dynamic programming (DP) procedure is applied to derive a temporal transformation function to align the time axis based on a STRAIGHT spectrogram within speech frequency bands.
A spectral warping function is defined as
\[
S'(\omega, t) = S(F^{-1}(\omega), t),
\]
\[
F(\omega) = \sum_{v=[a,i,u,e,i,o]} P(c(t) | \theta_v) F_v(\omega),
\]
\[
\alpha(F(\omega) - \omega) + \omega,
\]
where \(\omega\) is the frequency and \(F(\omega), F^{-1}(\omega)\) are a frequency warping function and its inverse function, respectively. A speaker dependent frequency warping function for each vowel \(v\) is \(F_v(\omega)\). These functions are automatically designed by a DP-base method [7] from the average spectrum calculated from one utterance consisting of consecutive vowels. \(\theta_v\) are the average spectra modeled as MFCCs that are also automatically calculated from the utterance. \(P(c(t) | \theta_v)\) is the posterior probability. The probability function is modeled by a multi-variable Gaussian model in the MFCC domain. The frequency domain transformation function for each frame is composed from the spectral warping functions based on segment similarity to each smoothed average spectrum. Fig. 2 shows the templates for frequency domain transformation. Vertical and horizontal axes show transformed and reference frequencies, respectively.

![Fig. 1 Outline of speech morphing based on bilateral voice style conversion](image)

The level compensation function is defined as
\[
S'(\omega, t) = \sum_{p=[a,i,u,e,i,o]} \exp\left(\alpha L_p(\omega) P(c(t) | \theta_p)\right) S(\omega, t),
\]
where the speaker dependent spectrum level compensation function is \(L_p(\omega)\). These functions are automatically calculated by subtracting the reference smoothed average spectrum from the target smoothed average spectrum represented by the lower order MFCC. The smoothed average spectrum is calculated from the same utterance used to design the spectral warping functions. Global spectral compensation is shown in Fig. 3. Vertical and horizontal axes show compensation level and a frequency, respectively.

Automatically designed spectral warping functions and the spectral warped results between two different speakers are displayed on the left and right panels in Fig. 2, respectively. On the right panel in Fig. 2, blue, red, and black lines show reference, converted, and conversion target spectra, respectively. Frequencies at the spectral peak of the blue line move to a suitable frequency (red line) of the black line by spectral warping. Level compensation functions are displayed in Fig. 3. The reference frequency at about 3 kHz corresponds to the warped frequency at about 2 kHz in frequency warping functions for vowel /o/. This over warping is over caused by the error of DP-base estimation. Spectral level is emphasized in lower frequency bands such as less than 1 kHz.
Figure 4 shows the average MFCC for each vowel. Posterior probabilities are calculated based on these MFCCs. Fig. 5 shows the posterior probability for such a sentence as ‘do da i’ (‘foundation’ in Japanese). Vertical and horizontal axes show posterior probability and time, respectively. The vertical black line in Fig. 5 shows manually assigned phonetic boundaries as a reference. Note that these boundaries are not required in the proposed procedure. Concerned vowel posterior probability is dominant in each vowel segment. No posterior probability is dominant in any non-vowel segment. These two features are important for the proposed procedure. Relying on posterior probability, spectral warping degree and level compensation degree are adaptively changed without phonetic boundary information and anchoring points. Voice style conversion based on vowel information is based on the following assumptions:

- The relevant coordinate mapping is continuous and monotonic both in the time and frequency domains. Its first order partial derivatives are also continuous.
- Relevant coordinate mapping is a direct product of the time domain coordinate mapping and frequency domain coordinate mapping.
- Deviations from the relevant coordinate mapping yield perceptually larger distortions for vowels than for consonants and transitional parts.

These assumptions suggest that a weighted average of template functions based on spectral proximity to the templates provides close approximation of the relevant coordinate transformation function in the vicinity of the templates. Regularly aligned dots and frequency warping images are applied to regularly aligned dots to automatically designed spectral warping functions and shown on the left and right panel in Fig. 6.

**Fig. 2 Spectral warping functions and warped spectrum for vowel /u/**

**Fig. 3 Automatic designed spectral level compensation functions**
Converted speech $A'$ is not entirely the same as conversion target speech $B$ because voice style is converted without information of the target itself. However, bilateral voice style conversion enables speech materials to be morphed. Morphed speech between bilateral voice style conversion speeches with appropriate weight are placed around the line $A$ to $B$ if
converted speech A’ and B’ do not greatly different from the real targets. This suggests the importance of improving style conversion procedure to improve the morphing procedure.

**EXPERIMENTS**

Preliminary subjective listening tests (ABX test) were conducted to evaluate the proposed method. The synthesized speech sounds morphed by the proposed method were compared to the synthesized speech sounds morphed by the original STRAIGHT-based morphing method, and the original STRAIGHT-based morphing method without frequency axis transformation. Six types of stimuli morphed between all possible pairs among 4 speakers at morphing ratio of 0.5 were prepared for each method. One sentence ‘ta no shi N de ru ?’ (“Are you having fun?”) was used for the test. The testee, one of the authors, evaluated a stimulus two times.

Schêve’s comparison between the proposed method and the original STRAIGHT-based morphing method without frequency alignment was conducted on a common time axis. Significant differences were recognized (p<0.014). This result show that automatic frequency warping and level compensation are limitedly effective to morph speech samples.

According to Schêve’s comparison, significant differences were recognized between the proposed method and the original STRAIGHT-based morphing method (p<0.001). Detecting was easy because the time axis is different between these methods. To focus on differences of spectral shape, Schêve’s comparison between the proposed method and the original STRAIGHT-based morphing method was conducted on a common time axis. Significant differences were recognized between the proposed method and the original STRAIGHT-based morphing method (p<0.001). These results show that synthesis parameters are insufficient to align on the time-frequency axis.

**CONCLUSIONS**

An automatic speech morphing method is proposed that eliminates the need for manual reference point assignment procedures in designing time-frequency coordinate mapping. Schêve’s comparison between the proposed method and the original STRAIGHT-based morphing method without frequency alignment was conducted on a common time axis. Significant differences were recognized (p<0.014). Based on Schêve’s comparison, significant differences were recognized between the proposed method and the original STRAIGHT-based morphing method (p<0.001). Although required to improve the proposed method, automatic frequency warping and level compensation are limitedly effective to morph speech samples.

**ACKNOWLEDGEMENT**

This research is partly supported by MEXT (Ministry of Education, Culture, Sports, Science and Technology) leading e-Society project and Basic Research Programs - CrestMuse project - conducted by JST (Japan Science and Technology Agency).

**References:**