Automatic assignment of anchoring points on vowel templates for defining correspondence between time-frequency representations of speech samples

Toru Takahashi, Masashi Nishi,
Toshio Irino, Hideki Kawahara

Faculty of Systems Engineering, Wakayama University, Japan
tall@sys.wakayama-u.ac.jp

Abstract

The automatic assignment of anchoring points is proposed to define the correspondence between the time-frequency representations of speech samples for speech morphing, speech texture mapping, and so on. The correspondence is modeled as a set of segmental bilinear function. These model parameters are called anchoring points. Although, the correspondence significantly affects the quality of such manipulated speech sounds as morphed and texture mapped speech sounds, anchoring points were manually aligned on time-frequency representations.

Anchoring points should be placed at auditorily important locations. When a spectrogram is presented as a time-frequency representation, auditorily important locations are given by formant frequencies around vowel transitions. The central idea of the proposed method is to prepare vowel template spectra with pre-assigned anchoring points in advance and to deform one of the templates to match the input speech spectrum. Finally, anchoring points on the input spectrum are copied from pre-assigned anchoring points.

Experimental results suggest that the naturalness of morphed speech based on the proposed automatic assignment method has equivalent quality to STRAIGHT synthetic speech samples.

Index Terms: template deformation, anchoring point, STRAIGHT, emotional speech synthesis, speech morphing

1. Introduction

Defining correspondence between one time-frequency representation and another plays an important role in speech morphing [1, 2], emotional style mapping [3], and speech texture mapping [4] procedures. Correspondence is explicitly represented as a mapping function. That can be designed by a 2-dimensional dynamic programing (2D DP) base method [5], but its computational cost is expensive to correspond between time-frequency representations. In this article, the mapping function is modeled as a segmental bilinear function defined by anchoring points. Since the mapping function is designed by fewer model parameters, computational cost decreases. This simplified model is reasonable for defining correspondence between time-frequency representations since some parameters are more auditorily important than others. In addition, the estimation accuracy of the parameters is more reliable than the other parameters. As alignment of anchoring points significantly affects the quality of manipulated speech sounds, anchoring points were manually aligned to auditorily important points in time-frequency representations. In some cases, the alignment procedure was repeated until the naturalness of the manipulated speech sounds achieved a desired quality. Manual alignment procedure was a huge obstacle for using speech morphing and speech texture mapping methods in various applications because it is time-consuming and tedious.

This article describes the methods that replace this process with an objective procedure. The central idea is to prepare vowel template spectra with pre-assigned anchoring points in advance and deform one of the templates to match the input speech spectrum. As the result of this deformation, pre-assigned anchoring points indicate auditory important points in the input speech spectrum. Anchoring points represent the locations on frequency and temporal coordinate system. The temporal coordinates of the anchoring points are determined by the phoneme labels annotated on the speech sample.

In the following section, first, parameter mapping based on anchoring points is described. Then morphed speech based on the automatic assignment of anchoring points is evaluated subjectively. Finally, this article is concluded.

2. Parameter mapping based on anchoring points

A mapping function represents a relationship of one parameter in a time-frequency representation to another. Figure 1 displays those parameters in each time-frequency coordinate system that are individually mapped into parameters in a common time-frequency coordinate system. Anchoring points in each coordinate system are drawn as circles. The mapping function is modeled as a segmental bilinear function defined by anchoring points. These anchoring points in a time-frequency representation are related to those in another time-frequency representation with a one-to-one-relation. Parameters in a quadrilateral area constructed from four anchoring points in an individual coordinate system are usually mapped to parameters in a square or rectangle area. Once parameters are represented in a common coordinate system, it is possible to carry out an operation between one time-frequency representation and another to modify the time-frequency representation. Finally, the modified parameters are mapped to a target time-frequency coordinate system.

2.1. Template spectrum with pre-assigned anchoring points

A template spectrum with pre-assigned anchoring points is formulated in this subsection. First, a vowel template spectrum is formulated. Then a method for aligning pre-assigned anchoring points to the template spectrum is described.

Template spectra are derived from speech samples in a database for each vowel: $/a/$, $/e/$, $/i/$, $/o/$, and $/u/$. For Japanese, a vowel template spectrum $\overline{S}^{/\xi/}(\omega)$, $(/\xi/ \in /a/, /e/, /i/, /o/, /u/)$ is defined as an average spectrum:

$$\overline{S}^{/\xi/}(\omega) = \frac{1}{T}\sum_{t\in L} S(t,\omega), \quad (L = \{t|label(t) \text{ is } /\xi/\}), \tag{1}$$

where spectral intensity at time t and frequency ω is denoted as $S(t,\omega)$. Function $label(t)$ returns a phonetic label at time t. The number of elements in set L is denoted as T.

Pre-assigned anchoring points are arranged where formants appear in the template spectrum, i.e., F1, F2, F3, and F4. Resonance frequencies calculated based on the AR model [8] are associated with spectral peak frequencies including formant frequencies. The AR model is powerful for estimating resonance frequencies; however it is a non-trivial problem to associate the resonances with F1, F2, F3, and F4. There are many reasons for this difficulty: 1) F1 is affected by glottal formants, 2) Two formants are merged and are apparently observed as one; 3) A reasonable number of AR model coefficients is unknown.

It is assumed that resonance frequencies, which are NOT associated with formant frequencies, are distributed randomly. A histogram of resonance frequencies shows local peaks associated with formant frequencies. Pre-assigned anchoring points are aligned to such frequencies. Based on this histogram approach, the results of aligning to pre-assigned anchoring points in the template spectrum are shown in Fig. 2.

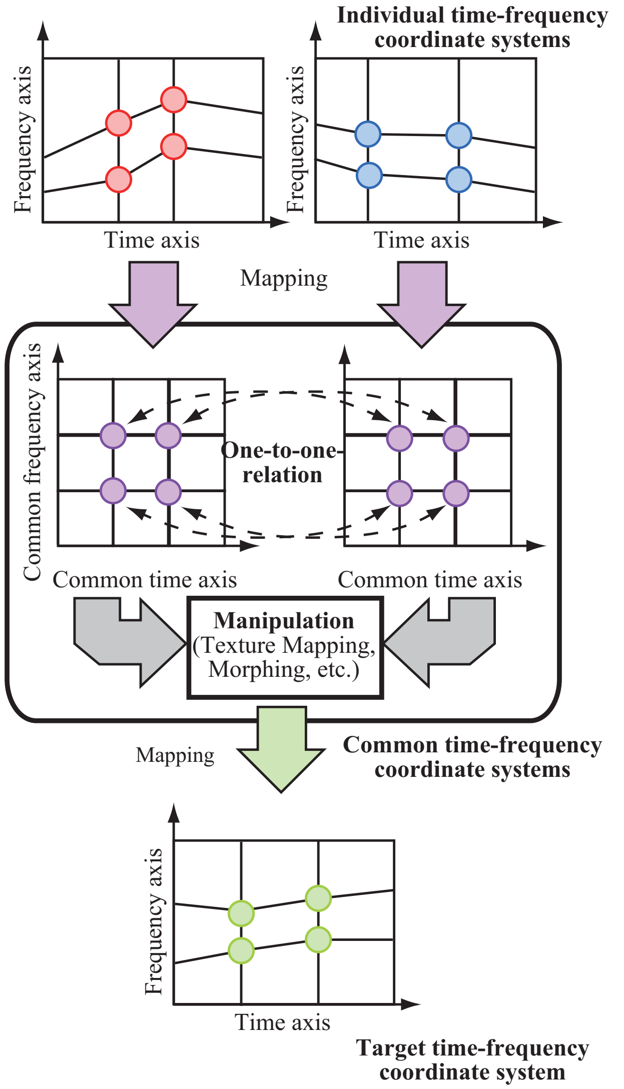

Figure 1: *Time-frequency parameter mapping with anchoring points*

2.2. Automatic anchoring point alignment algorithm

The automatic alignment procedure consists of two steps. First, an input spectrum is matched against a template spectrum with pre-assigned anchoring points. Spectral distance D between input spectrum $S(\omega)$ and deformed template spectra $\overline{S}^{/\xi/}(f(\omega))$ is minimized with respect to all possible deformations. The cost function is defined as follows:

$$f_{min} = \arg\min_{f} D, \tag{2}$$

$$D = d(\overline{S}^{/\xi/}(\omega), \overline{S}^{/\xi/}(f(\omega))), \tag{3}$$

where $f(\omega)$ is a frequency warping function for the deforming spectrum. A frequency warping function is defined as a

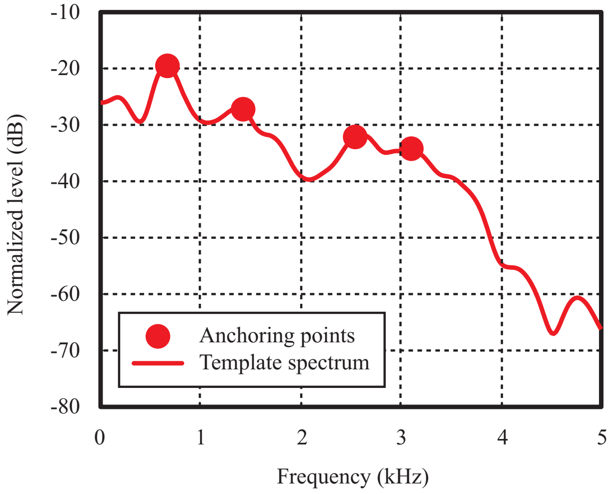

Figure 2: *Template spectrum for vowel /a/ with anchoring points*

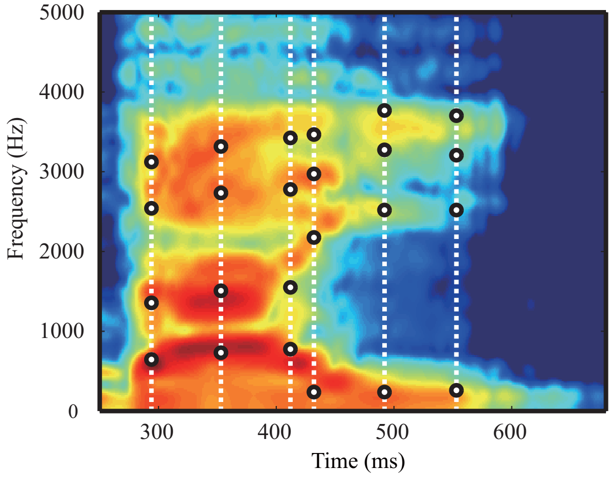

Figure 3: *Automatic aligned anchoring points*

segmental bilinear function:

$$f(\omega) = \frac{\tilde{\omega}_n - \tilde{\omega}_{n-1}}{\omega_n - \omega_{n-1}}\omega + \frac{\omega_{n-1}\tilde{\omega}_n - \tilde{\omega}_{n-1}\omega_n}{\omega_{n-1} - \omega_n}, \qquad (4)$$

where the frequencies of pre-assigned anchoring points are denoted as $\omega_n (n = 0, \ldots, N-1, \omega_n < \omega_{n+1})$, and the deformed frequencies of pre-assigned anchoring points are denoted as $\tilde{\omega}_n (n = 0, \ldots, N-1, \tilde{\omega}_n < \tilde{\omega}_{n+1})$.

Second, anchoring points in the input spectrum are copied from pre-assigned anchoring points. The coordinates of the anchoring points are identical to the pre-assigned anchoring points in the deformed template spectrum. Finally, the frequencies of the anchoring points are represented as $f_{min}(\omega_n), (n = 0, \ldots, N)$. Automatic aligned anchoring points are displayed in Fig. 3.

3. Discussion

A subjective experiment confirms that morphed speech using automatically assigned anchoring points is natural enough.

3.1. Speech materials

The speech samples used for the subjective tests were morphed from recorded speech samples in an emotional speech database. That consisted of speech samples narrated by four professional actors. The three Japanese sentences evaluated by testees are shown in Table 1. Five types of styled speeches, 'Neutral,' 'Bright,' 'Excited,' 'Angry,' and 'Furious,' were sampled at 44.1 kHz with a 16 bit linear PCM format in an acoustical room under background noise of 30 dB. An omnidirectional microphone, WM-8160, Matsushita Electric Industrial Co., Ltd. was used for recording. The microphone was set 15 cm from the speaker.

3.2. Evaluation

To evaluate the effectiveness of the automatically aligned anchoring points, the morphed speech was tested subjectively at morphing ratios of 0.0, 0.25, 0.5, 0.75, and 1.0. Morphed speech at morphing ratios of 0.0 and 1.0 were equivalent to reference and target speech, respectively. Three pairs of reference and target styles, i.e., 'Neutral-Angry,' 'Neutral-Bright,' and 'Neutral-Excited' were used for evaluations. 180 kinds of stimuli were presented (four speakers, three styles, five morphing ratio steps, and three kinds of sentences). Each kind of stimuli was presented twice per test. All stimuli were presented in random order.

The ten testees, confirmed to possess normal auditory capacity, were university students (males: 6, females: 4). In the evaluation procedure, speech stimuli were presented in diotic conditions through headphones, SNNHEISER HD-580. The equipment for presenting speech stimuli was calibrated by using HATS, B&K 4128.

Testees were instructed to evaluate the naturalness of speech samples and to grade them on five levels from very unnatural (1) to very natural (5). Naturalness was defined as a quality of human-like speech, not as an expressed speech style.

Figures 4 and 5 show the average rates over all testees for sentences 1 and 2, respectively. Morphed speech sounds at morphing ratios of 0.0 and 1.0 are equivalent to the analysis-by-synthesis speech sounds of reference and

Table 1: *Test sentences*

#	Japanese	Translation
1	kaitoudekimasita	Thawing is finished.
2	oyugawakimasita	The Water has boiled.
3	kansouwohajimemasu	Drying has started.

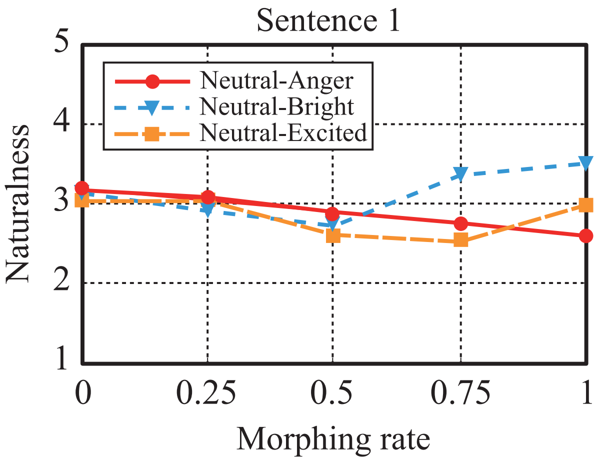

Figure 4: *Naturalness of morphed speech: Sentence 1*

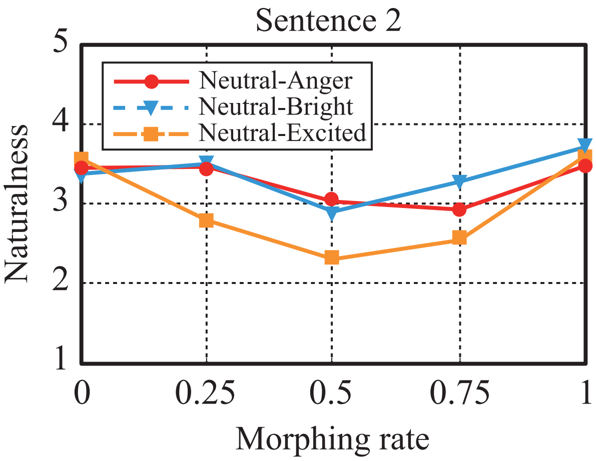

Figure 5: *Naturalness of morphed speech: Sentence 2*

target speech sounds. As other morphed speech sounds are synthesized with a parameter modification process, it was anticipated that the average rates at morphing ratios of 0 and 1 would always be higher than morphed speech sounds at other morphing ratios. Fig. 5 shows the anticipated trend, which differs from 4.

Although singing cannot be identified by speech sounds, this trend is found in morphed singing sounds, as mentioned in [9]. These results support that morphed speech based on well-aligned anchoring points has the same naturalness as synthetic speech based on STRAIGHT.

The naturalness of almost all morphed speech is better than the synthetic anger style speech (In Fig. 4, the morphing rate was 1.0). At a 5% level of significance [7], morphed speech has the same naturalness as synthetic speech based on STRAIGHT. Therefore, experimental results suggest that the proposed alignment method is adequate to align anchoring points for defining correspondence between time-frequency representations of speech samples.

4. Conclusion

The automatic assignment of anchoring points was proposed for defining correspondence between the time-frequency representations of speech samples. Since anchoring points were manually assigned, assignment was a huge obstacle for using speech morphing and speech texture mapping methods in various applications. However, the proposed method eliminates the obstacle from the anchoring points assignment process. Subjective experiments suggest that the proposed automatic assignment method maintains the neutralness of morphed speech sounds and is adequate to assign anchoring points.

5. Acknowledgments

This research was partly supported by the Ministry Education, Culture, Sports, Science and Technology leading e-Society project.

6. References

[1] M.Slaney, M.Covell, and B.Lassiter, ”Automatic audio morphing,” Proc. IEEE ICASSP 1996, Vol. II, pp. 1001-1004, 1996.

[2] H.Kawahara and H.Matsui, ”Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation,” Proc. IEEE ICASSP 2003, Vol. I, pp. 256–259, 2003.

[3] T.Takahashi, T.Fujii, M.Nishi, H.Banno, T.Irino, and H.Kawahara, ”Voice and Emotional Expression Transformation based on Statistics of Vowel Parameters in an Emotional Speech Database,” Proc. Interspeech 2005, pp. 537–540, Sept., 2005.

[4] T.Takahashi, T.Irino, and H.Kawahara, ”General framework for flexible speech style manipulation and synthesis,” Proc. WESPAC IX 2006, June, 2006.

[5] S.Uchida and H.Sakoe, ”Monotonic and Continuous Two-Dimensional Warping Based on Dynamic Programming,” Institute of Electronics, Information and Communication Engineers, Vol. J81-D-II, No. 6, pp. 1251–1258, 1998. (In Japanese).

[6] H.Matsui and H.Kawahara, ”Investigation of Emotionally Morphed Speech Perception and its Structure Using a High Quality Speech Manipulation System,” Proc. EUROSPEECH 2003, pp. 2113–2116, Sept., 2003.

[7] G.A.Ferguson and Y.Takane, ”Statistical analysis in psychology and education,” McGRAW-HILL, Sixth edition. 1989.

[8] F.Itakura and S.Saito, ”Analysis synthesis telephony based on the maximum likelihood,” Proc. 6th International Congress on Acoustic, 1968.

[9] T.Yonezawa, N.Suzuki, K.Mase, and K.Kogure, ”Gradually changing expression of singing voice based on morphing,” Proc. Interspeech 2005, pp. 541–544, Sept., 2005.