AUDITORY MORPHING BASED ON AN ELASTIC PERCEPTUAL DISTANCE METRIC IN AN INTERFERENCE-FREE TIME-FREQUENCY REPRESENTATION

Hideki Kawahara† *and Hisami Matsui*

Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Japan
† kawahara@sys.wakayama-u.ac.jp

ABSTRACT

An elastic spectral distance measure based on a F0 adaptive pitch synchronous spectral estimation and selective elimination of periodicity interferences, that was developed for a high-quality speech modification procedure STRAIGHT [1], is introduced to provide a basis for auditory morphing. The proposed measure is implemented on a low dimensional piecewise bilinear time-frequency mapping between the target and the original speech representations. A preliminary test results of morphing emotional speech samples indicated that proposed procedure provides perceptually monotonic and high-quality interpolation and extrapolation of CD quality speech samples.

1. INTRODUCTION

A F0 (fundamental frequency) adaptive pitch synchronous time-frequency representation [1] was introduced to enable versatile speech manipulations, for example auditory morphing [2, 3], while keeping manipulated speech quality virtually indistinguishable from natural ones. Auditory morphing has a lot of potential applications in post processing multimedia and musical contents. It also provides a powerful research tool for speech perception, especially for non-linguistic and para-linguistic information. For example, control of emotional aspect of synthetic speech, that is a hot topic in human computer interaction, needs effective understanding of physical correlates of that information. There are several review papers on emotional speech [4, 5] However, no effective and high-quality control method was established yet, partly because of methodological limitations as well as motivations of research. Auditory morphing enables an exemplar-based strategy, in addition to conventional research strategies (for example, analytical and synthetic) for investigating this hard problem.

Because of its straightforward speech information representations, STRAIGHT [1] has been claimed to be the best-fit method for versatile speech manipulations. Morphing is one of such manipulations. This article introduces a simple implementation of a speech morphing procedure to substantiate this claim. It is worthwhile to note that the proposed morphing method here is not automatic, unlike other morphing literatures [2, 3]. It involves manual placement of anchor points in the reference and the target speech representations. However, this seemingly "low-tech" feature of the proposed method allows researches full control of parameter mapping between instances. This is a desirable feature for a research tool. It should be noted that the assignment of anchor points is done on the interference free time-frequency representation [1], which is close to familiar spectrograms. It is also important to note that careful manual manipulation using STRAIGHT-based method makes it possible to produce a highly realistic morphed speech. This is very important for a research strategy called "systematic downgrading" described bellow.

2. SYSTEMATIC DOWNGRADING

A strategy called systematic downgrading was initially proposed in the context of research on scat singing [6], where non-linguistic and pare-linguistic information plays indispensable roles. The central idea of "systematic downgrading" is to keep test stimuli as ecologically relevant (in another word, highly natural) as possible. It is important to use ecologically relevant stimuli, because human perceptual systems can be highly nonlinear, meaning it is generally difficult to draw dependable conclusions for human responses to highly complex signals (for example speech) only based on responses to elementary stimuli such as tone, tone bursts, clicks and noise. It is also important to have means to manipulate physical parameters of the stimuli in a well-defined manner. The STRAIGHT-based morphing fulfils requirements on ecological relevance (high quality resynthesized speech) and precise control of physical parameters simultaneously.

The following steps outline "systematic downgrading" in case of investigating regularities in emotional control.

†Also an invited researcher of ATR and the principal investigator of CREST's "Auditory Brain Project".

(1) Prepare the reference speech and the target speech with intended emotions.
(2) Morph the reference speech to the target speech by careful manual transformation of parameters.
(3) Extract regularities in the manual transformation and design series of approximation functions of the transformation.
(4) Morph the reference speech by the approximation functions AND refine it with additional manual modifications.
(5) Repeat step (3) and (4) until satisfactory approximation function is designed.

The procedure is a generalized version of the "null point procedure", which is a common practice to minimize disturbances to the measured system. It keeps the critical subjective evaluation to be performed only for high-quality (ecologically relevant) stimuli.

3. AUDITORY MORPHING

Morphing is a procedure to regenerate a signal from a representation on a shortest trajectory between anchor points in an abstract distance space with a distance metric d_{fx}. When there is no ambiguity, it is also possible to extrapolate the shortest trajectory outside the anchor points.

It is necessary to introduce an approximation that yields practical implementation of this general morphing procedure. One such approximation for speech morphing is to define the new distance ${d_{cp}}^1$ as a composite operations of a coordinate transformation $\mathcal{T}$ and a localized distance metric d_{pp}.

$$d_{fx} \simeq d_{cp} = d_{pp}(s_{ref}(\lambda, \tau)), s_{tgt}(\mathcal{T}(\lambda, \tau))) \qquad (1)$$

where ref and tgt represents "reference" and "target" respectively. If the transformation $\mathcal{T}$ does not have any penalty due to the transformation, and if the localized distance metric is Euclidean, the morphing procedure is reduced to a linear interpolation on representations represented on the reference coordinate. The proposed procedure described below is based on this approximation.

There are several technical issues to implement the procedure. Specifically, the coordinate system and the localized distance metric must reflect auditory perceptual characteristics, and the transformation must be as simple as possible. In this article, the time-frequency plane is used as the coordinate system. The transformation is represented as a simple piecewise bilinear transformation, because, unlike the image morphing, the time-frequency coordinate is not isotropic. Practically, only up to 5 anchor points on a frequency axis at one temporal location and up to 4 temporal anchor points for one CV syllable are found sufficient. For the fundamental frequency, it is relevant to morph the parameter in the log-frequency domain, because the F0 dynamics is represented in terms of a linear dynamical equation in the log-frequency domain [7]. For the spectral density, morphing is calculated on dB representation, because it is one of relevant approximations of intensity perception. The time-frequency periodicity index [8, 9] is also transformed by the same mapping function.

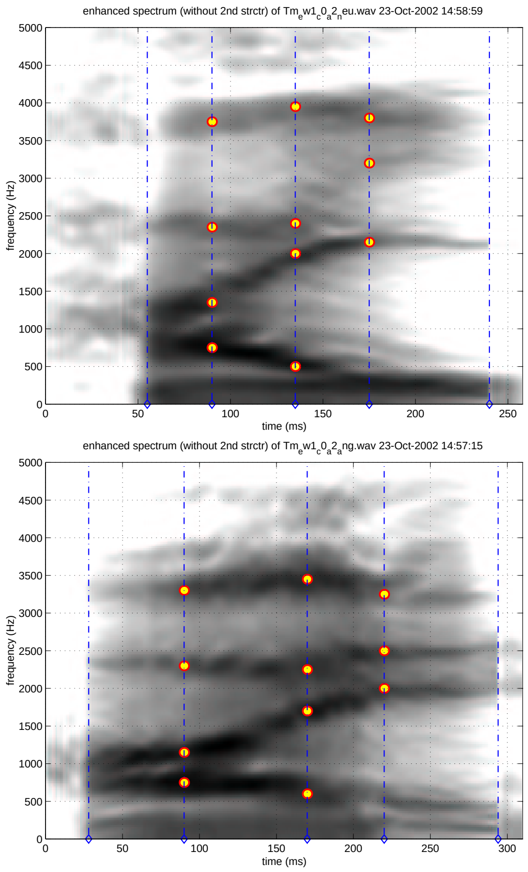

Fig. 1. Smooth spectrographic representations of words played by a male actor under neutral (upper) and angry (lower) emotional conditions. Anchor points in the time-frequency domain are plotted as open circles and temporal anchors are plotted as vertical dash-dot lines. (Lower frequency portion (≤ 5000 Hz) of time-frequency representations are shown to clarify contrasts.)

4. MORPHING EXAMPLE

This section illustrates the procedure by a step-by-step example of morphing emotional speech materials, which will

[1] Strictly speaking, this approximation does not define distance metric. The approximation does not satisfy the requirement for distance metric; $d_{cp}(a, b) = d_{cp}(b, a)$. However, until it is inevitable, this simplified definition is used in this article.

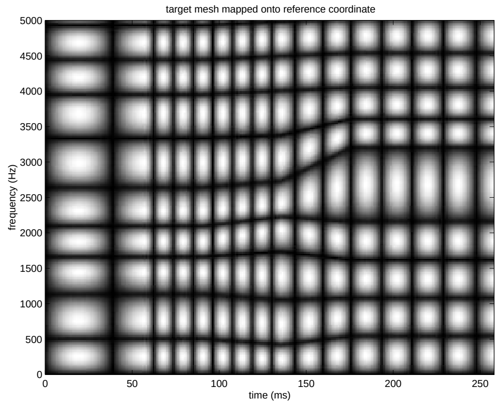

Fig. 2. Regular time-frequency grid in the target coordinate transformed into the reference coordinate.

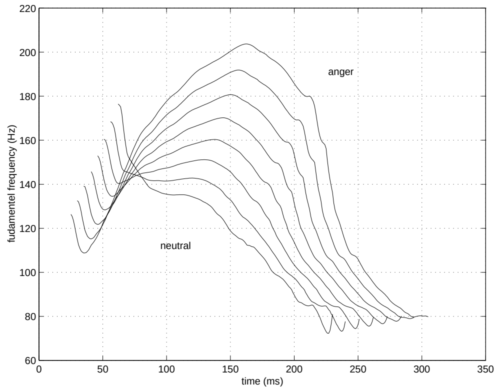

Fig. 3. Morphed F0 patterns spanning from -0.2 to 1.2 with step size 0.2 in morphing rate.

be described in Section 5.

The first step is to analyze speech parameters. STRAIGHT analysis produces a time-frequency (spectrographic) representation that does not have interference due to periodicity, F0 trajectory and a periodicity map. Figure 1 shows the time-frequency representations of words spoken by a professional male actor under two different emotional conditions (neutral and anger).

The next step is to assign anchor points to define correspondence between two representations. Based on visual inspections and basic phonological knowledge, the anchor points in the time-frequency domain were selected as shown in the same figure (Figure 1).

The third step is to design a piecewise bilinear transformation to map the target time-frequency coordinate onto the reference coordinate using the information given as anchor points. Figure 2 shows how a regular time-frequency grid is transformed by the mapping designed by this step.

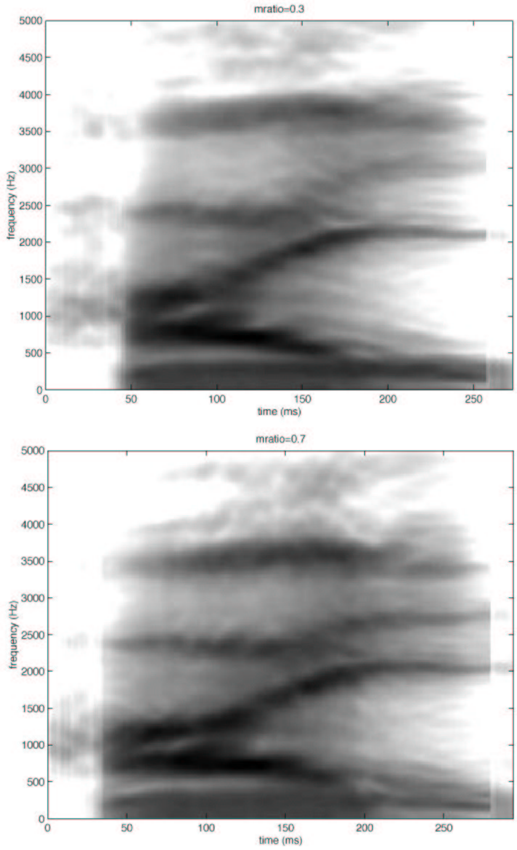

Fig. 4. Morphed time-frequency representations between "neutral" and "angry" conditions. (morphing rate: 0.3 (upper) and 0.7 (lower)).

The fourth step is to calculate morphed parameters in the reference coordinate system. All information necessary for reproduce speech sounds were represented in the reference coordinate system after the first three steps. Let a symbol Θ_S^C to represent a set of parameters spoken (or generated) at the condition S and represented in the coordinate system C. Then, the morphed representation with a morphing rate r is calculated by the following operation.

$$\Theta_{mrp}^{ref}(r) = (1-r)\Theta_{ref}^{ref} + r\Theta_{tgt}^{ref} \qquad (2)$$

where mrp represents morphing.

The fifth step is simply a backward transformation of the morphed parameters into the morphed coordinate system. It is important to note that the transforming operation can be separable into two successive steps; the frequency transformation and the temporal transformation. Figure 3 shows morphed F0 trajectories with morphing rates ranging from -0.2 to 1.2 with 0.2 steps. Figure 4 also shows morphed time-frequency representations for morphing rates 0.3 and 0.7. The final step is to resynthesize speech using these morphed parameters.

4.1. Extension to non stationary morphing rates

As the transformation is represented using a separable bilinear transform, a temporarily non stational morphing rate $r(t)$ is easily implemented using the following temporal mapping $\mathcal{T}_{mrp}(t;r)$.

$$\mathcal{T}_{mrp}(t;r) = \int_{t_0}^{t} \left((1 - r(\tau)) + r(\tau)\frac{\partial \mathcal{T}^{-1}}{\partial \tau} \right) d\tau \qquad (3)$$

This extension only requires a modification in the fourth and the fifth steps to use a time dependent morphing rate $r(t)$ instead of using a constant r.

5. PRELIMINARY TEST RESULTS

A portrayal emotional speech database, which was recorded using professional (one male and one female) actors and a recording studio for professional use, was designed for development and evaluation of morphing functions. Seven basic emotional expressions (neutral, angry, sadness, happiness, fear, surprize and disgust) were prepared for three target Japanese words (/hai/, /iie/ and /koNnitiwa/: “yes”, “no” and “hello” in English, respectively). Target words were recorded twice under four different contextual conditions (preceding, following and surrounding carrier sentences and isolated pronunciation), two sentence types (declarative and interrogative). Speech sounds captured by an omni-directional condenser microphone (Sanken CU-41) and amplified were directly sampled at 44100 Hz and digitized in 16 bits.

Multi dimensional analyses on subjective scoring results of perceived emotional attributes were conducted to select the most salient utterance for preliminary tests. Four (two male and two female) subjects participated in this screening test. The word /hai/ spoken by a male actor under angry condition was selected as the most salient emotional speech. A stimulus continuum ranging -0.2 to 1.2 in 0.1 steps in morphing rate was generated using angry and neutral utterances to find the JND (just noticeable difference) of morphing rate. The JND was about 0.25 in this case.

Stimulus continua for all combinations of seven emotional conditions were generated using -0.25 to 1.25 in 0.25 steps in morphing rates. Informal subjective tests indicated that each continuum was perceived monotonic and high-quality. A comprehensive set of subjective tests is currently in progress.

Examples of resynthesized speech using STRAIGHT-based morphing can be found in the following URL.
`http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTdemo/`

6. SUMMARY

A procedure for auditory morphing based on a high-quality speech manipulation system STRAIGHT [1] is introduced. This implementation enables flexible morphing in the whole speech parametric domain. A set of preliminary tests were conducted using an emotional speech database, which was recorded by using professional actors. The test results of morphing emotional speech samples indicated that proposed procedure provides perceptually monotonic and high-quality interpolation and extrapolation of speech samples.

Acknowledgement This work was partly supported by CREST of JST and Grant-in-Aid for Scientific Research (A) 12040241 and (C) 14380165.

7. REFERENCES

[1] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,” *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[2] M. Slaney and B. Lassiter, “Automatic auditory morphing,” in *Proc. ICASSP’96*, 1996, vol. 2, pp. 1001–1004.

[3] H. Banno, K. Takeda, K. Shikano, and F. Itakura, “Speech morphing by independent interpolation of speech envelope and source excitation,” *J. of IEICEJ*, vol. J81-A, no. 2, pp. 261–268, 1998, [In Japanese].

[4] I.R. Murray and J.L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[5] M. Schroeder, “Emotional speech synthesis: A review,” in *Proc. EUROSPEECH Scandinavia*, 2001, pp. 561–564.

[6] H. Kawahara and H. Katayose, “Scat generation research program based on straight, a high-quality speech analysis, modification and synthesis system,” *J. of IPSJ*, vol. 43, no. 2, pp. 208–218, 2002, [in Japanese].

[7] H. Fujisaki, “A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour,” in *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, O. Fujimura, Ed., New York, 1998, pp. 347–355, Raven Press.

[8] Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigné, and Roy D. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity,” in *Proc. Eurospeech’99*, 1999, vol. 6, pp. 2781–2784.

[9] Hideki Kawahara, Jo Estill, and Osamu Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight,” in *Proc. 2nd MAVEBA*, Firenze, Italy, 2001, [CD ROM].