TEMPORALLY VARIABLE MULTI-ASPECT AUDITORY MORPHING ENABLING EXTRAPOLATION WITHOUT OBJECTIVE AND PERCEPTUAL BREAKDOWN

*H. Kawahara, R. Nisimura, T. Irino**

Design Information Sciences Department
Faculty of Systems Eng., Wakayama Univ.
930 Sakaedani, Wakayama, 640-8510, Japan

[1]*M. Morise,* [2]*T. Takahashi,* [3]*H. Banno*

[1]Kwansei Gakuin University, Japan
[2]Kyoto University, Japan
[3]Meijo University, Japan

ABSTRACT

A generalized framework of auditory morphing based on the speech analysis, modification and resynthesis system STRAIGHT is proposed that enables each morphing rate of representational aspects to be a function of time, including the temporal axis itself. Two types of algorithms were derived: an incremental algorithm for real-time manipulation of morphing rates and a batch processing algorithm for off-line post-production applications. By defining morphing in terms of the derivative of mapping functions in the logarithmic domain, breakdown of morphing resynthesis found in the previous formulation in the case of extrapolations was eliminated. A method to alleviate perceptual defects in extrapolation is also introduced.

***Index Terms*—** speech analysis, speech processing, speech synthesis, auditory system, computer music

1. INTRODUCTION

The auditory morphing procedure [1] based on STRAIGHT [2] is widely applied in scientific research [3], interface design [4] and singing manipulations [5] because of its highly natural manipulated sounds and controllability of precise physical parameters. However, when the manipulation is extended into the region where extrapolation of one or more aspects of representational parameters is needed, reproduced sound quality deteriorates very rapidly. Moreover, sometimes the morphing algorithm fails because monotonicity of mapping is violated by extrapolation. This is one of the issues to be solved in this paper.

The other issue addressed in this paper is extension to temporally variable morphing rates. Requests for such extension were raised in various applications. Real-time manipulation of morphing rates for singing performance is one such example. It requires formulation in terms of an incremental algorithm. In speech perception research, careful preparation of test stimuli to investigate contributing factors in a specific perceptual aspect has to be done in advance. It requires formulation in terms of a batch processing algorithm.

In addition to these motivations, recent introduction of a new STRAIGHT called TANDEM-STRAIGHT [6] made reformulation of the morphing procedure crucial because TANDEM-STRAIGHT significantly reduced computational cost and demand on memory space. Introduction of TANDEM-STRAIGHT made the morphing procedure the limiting factor of the total throughput and performance in manipulating speech materials. Taking these requests into account, the proposed method is implemented using "just-in-time" program architecture.

*Partially supported by Grants-in-Aid for Scientific Research (A) 19200017 by JSPS and the CrestMuse project by JST.

2. BACKGROUND: STRAIGHT-BASED MORPHING

The morphing procedure based on STRAIGHT in our previous proposal is briefly introduced to provide background for the newly proposed reformulation.

STRAIGHT (and TANDEM-STRAIGHT as well) decomposes input speech in terms of fundamental frequency (F0), aperiodicity spectrogram and interference-free spectrographic representation (STRAIGHT spectrogram). They are represented as a function of time or time and frequency. An utterance is considered to be one point in an abstract high-dimensional space spanned by these five representational aspects (F0, aperiodicity and STRAIGHT spectrogram morphing rates and morphing rates for time axis conversion and frequency axis conversion). The auditory morphing is a procedure to draw a trajectory from one point to the other in this abstract space.

The original morphing algorithm [1] was implemented as a five-stage procedure. The first step is to extract parameters (F0, aperiodicity and STRAIGHT spectrogram) of each utterance. The second step is to align time-frequency coordinates of parameters of two exemplar utterances. The third step is to interpolate (and if necessary extrapolate) parameters represented on the aligned time-frequency coordinates according to the given morphing rate(s). The fourth step is to deform the time-frequency coordinates according to the given morphing rate(s). The final step is to resynthesize sound using the morphed parameters on the morphed time-frequency coordinate.

The time-frequency conversion function to align and deform the time-frequency coordinate is designed based on the time-frequency anchoring points assigned manually by the users. The conversion function is implemented as a collection of bilinear transformations of each time-frequency patch. A coordinate transformation T_{Am} from the coordinate of the example A to the morphed coordinate (denoted by m) is represented as follows:

$$T_{Am}(x) = (1-r)T_{AA}(x) + rT_{AB}(x)\ , \tag{1}$$

where x represents a value on the coordinate of speaker A, and T_{AA} and T_{AB} represent the transformation from and to the same speaker A and the transformation from speaker A to speaker B, respectively. The parameter r represents the morphing rate. When $r = 0$, it yields $T_{Am} = T_{AA}$ and is an identity mapping. Similarly, a set of parameters Θ on the aligned time-frequency coordinate is morphed using the following equation:

$$\Theta_m = (1-r)\Theta_A + r\Theta_B\ , \tag{2}$$

where $\Theta_m, \Theta_A, \Theta_B$ represent parameters of the morphed one, speaker A and speaker B respectively.

2.1. Failure in extrapolative morphing

This formulation works fine for interpolative morphing, where r resides inside the region $(0, 1)$. However, when $r < 0$ or $r > 1$, monotonicity of the mapping T_{Am} cannot be assured and results in failure of the morphing procedure because the inverse function of T_{Am} cannot be properly defined due to ambiguity in inversion. This is an objective breakdown in extrapolation.

Parameter conversion based on Eq. 2 also introduces defects in reproduced speech quality in the case of extrapolative morphing. A sharp spectral dip in one of the speakers yields a sharp spectral peak when extrapolation is introduced. Such a peak is perceived as annoying artificial timbre and results in severe quality degradation. This is perceptual breakdown in extrapolation. These objective and perceptual breakdowns form the first issue to be solved.

2.2. Extension to temporally variable multi-aspect morphing

The morphing rate r in Eqs. 1 and 2 was originally defined as a scalar value. It was extended to become a vector parameter $\boldsymbol{r}$ to define the morphing rate of each aspect independently. The next extension is to allow each morphing rate to be a function of time to meet demands described in the introduction.

3. MORPHING OF LOGARITHMIC DERIVATIVE

Positive parameters can be manipulated in terms of their logarithmic converted versions followed by exponential conversion to assure positivity of the resultant parameters. This common practice is applied to replace Eq. 1 for coordinate morphing.

Instead of using Eq. 1, the coordinate transformation T_{Am} is defined by the following equation and reduced into a simpler form:

$$\begin{aligned} T_{Am}(x_A) &= \int_0^{x_A} \exp\left(\log\left(\frac{dT_{Am}(\lambda)}{d\lambda}\right)\right) d\lambda \\ &= \int_0^{x_A} \exp\left((1-r_{AB})\log\left(\frac{dT_{AA}(\lambda)}{d\lambda}\right)\right. \\ &\quad \left. + r_{AB}\log\left(\frac{dT_{AB}(\lambda)}{d\lambda}\right)\right) d\lambda \\ &= \int_0^{x_A} \left(\frac{dT_{AB}(\lambda)}{d\lambda}\right)^{r_{AB}} d\lambda\,, \end{aligned} \tag{3}$$

because logarithmic conversion of the identity mapping vanishes. Suffixes of x and r explicitly represent associated coordinate systems. This formulation assures monotonicity of T_{Am} if the coordinate conversion T_{AB} from speaker A to B is monotonic. Similarly, the following related conversion functions are also monotonic under the same condition:

$$T_{mA}(x_m) = \int_0^{x_m} \left(\frac{dT_{AB}(\lambda)}{d\lambda}\right)^{-r_{AB}} d\lambda, \tag{4}$$

$$T_{BA}(x_B) = \int_0^{x_B} \left(\frac{dT_{BA}(\lambda)}{d\lambda}\right)^{r_{BA}} d\lambda, \tag{5}$$

$$T_{mB}(x_m) = \int_0^{x_m} \left(\frac{dT_{BA}(\lambda)}{d\lambda}\right)^{-r_{BA}} d\lambda. \tag{6}$$

These formulations are applicable to the temporal axis conversion as well as the frequency axis conversion. These new definitions completely eliminate the objective breakdown in extrapolation because monotonicity always holds. A solution to the perceptual breakdown in extrapolation is outlined in the section for discussion.

4. TEMPORALLY VARIABLE MORPHING RATES

The formulation introduced in the previous section is applied to extend morphing parameters to vary temporally. There are two alternatives for how to define temporally variable morphing rates, especially the temporally variable morphing rate of the time axis itself.

One possibility is to use the time axis in the real world, in other words, the time axis of the synthesized signal. This is suitable for real-time applications. The other possibility is to use a pre-deterimined morphed time axis, for example a morphed time axis using a temporally constant morphing rate. This is suitable for post-production applications and preparation of test stimuli of perceptual experiments.

4.1. Morphing rate definition on the "real" time axis

Assume that the morphing rate of the temporal axis $r_{AB}^{(t)}(t_s)$ is defined in terms of the time axis of the synthesized signal, in other words, the time axis of the real world. The suffix s is after "s" in "synthesis." The symbol $^{(t)}$ explicitly represents that the morphing rate is for the time axis. The following set of equations define an incremental algorithm for morphing with temporally variable morphing rates:

$$t_s = \int_0^{t_s} d\lambda, \tag{7}$$

$$T_{sA}(t_s) = \int_0^{t_s} \left(\frac{dT_{AB}(T_{sA}(\lambda))}{d\lambda}\right)^{-r_{AB}^{(t)}(\lambda)} d\lambda, \tag{8}$$

$$T_{sB}(t_s) = \int_0^{t_s} \left(\frac{dT_{BA}(T_{sB}(\lambda))}{d\lambda}\right)^{(r_{AB}^{(t)}(\lambda)-1)} d\lambda, \tag{9}$$

where the derivatives are used to update their time axes $t_A = T_{sA}(t_s)$ and $t_B = T_{sB}(t_s)$ and are calculated using these updated temporal coordinates. The morphed parameter set $\Theta_m(t_s)$ on the "real" time axis t_s is calculated using the following equation:

$$\begin{aligned} \Theta_m(t_s) &= (1-\boldsymbol{r}_{AB}(t_s))\Theta_A(T_{sA}(t_s)) \\ &\quad + \boldsymbol{r}_{AB}(t_s)\Theta_B(T_{sB}(t_s)). \end{aligned} \tag{10}$$

This is an incremental algorithm. It is suitable for real-time interactive applications, where morphing rates are continuously updated through interactions with the player/performer. Since STRAIGHT (TANDEM-STRAIGHT as well) uses morphed parameters at sparsely located discrete instances where excitation pulses are located, Eq. 10 needs to be calculated only when it is required. In other words, they are to be calculated at "just-in-time" when they are needed.

4.2. Morphing rate definition on a pre-defined time axis

In this case, morphing rates are defined on a time axis prepared in advance. Any time axis can be used if it has monotonic mapping onto both time axes of speakers A and B. Practically, one of the speakers' time axes may be preferably selected. A morphed time axis using a temporally constant morphing rate $r_{rAB}^{(t)}$ generalizes the last case by selecting the proper value of $r_{rAB}^{(t)}$.

Let t_r represent the pre-defined time axis (the reference time axis) using $r_{rAB}^{(t)}$ for the temporally constant morphing rate. Let T_{Ar} represent conversion from the time axis of speaker A to the reference time axis. Associated conversions are given by the following

equations:

$$t_r = T_{Ar}(t_A) = \int_0^{t_A} \left(\frac{dT_{AB}}{d\lambda}\right)^{r_{r_{AB}}^{(t)}} d\lambda, \quad (11)$$

$$T_{rA}(t_r) = \int_0^{t_r} \left(\frac{dT_{AB}}{d\lambda}\right)^{-r_{r_{AB}}^{(t)}} d\lambda, \quad (12)$$

$$T_{rB}(t_r) = \int_0^{t_r} \left(\frac{dT_{BA}}{d\lambda}\right)^{(r_{r_{AB}}^{(t)}-1)} dt, \quad (13)$$

where $T_{rA}(t_r)$ represents the transformation from the reference time axis to the time axis of speaker A and $T_{rB}(t_r)$ represents the similar transformation for speaker B.

Let $r_{AB}^{(t)}(t_r)$ represent the temporally variable morphing rate of the time axis defined on the reference time axis. Then, the time axis t_s of the morphed output speech signal and associated transformations are given by the following equations:

$$T_{rs}(t_r) = \int_0^{t_r} \left(\frac{dT_{AB}}{d\lambda}\right)^{(r_{AB}^{(t)}(\lambda)-r_{r_{AB}}^{(t)})} d\lambda, \quad (14)$$

$$T_{sr}(t_s) = \int_0^{t_s} \left(\frac{dT_{AB}}{d\lambda}\right)^{(r_{r_{AB}}^{(t)}-r_{AB}^{(t)}(\lambda))} d\lambda, \quad (15)$$

where $t_s = T_{rs}(t_r)$ and $t_r = T_{sr}(t_s)$.

The morphed set of parameters $\Theta_m(t_s)$ on the "real" time axis yielded by a set of temporally variable morphing rates $\boldsymbol{r}_{AB}(t_r)$ defined on the reference time axis is calculated using the following equation:

$$\Theta_m(t_s) = (1-\boldsymbol{r}_{AB}(T_{sr}(t_s)))\Theta_A(T_{rA}(T_{sr}(t_s))) + \boldsymbol{r}_{AB}(T_{sr}(t_s))\Theta_B(T_{rB}(T_{sr}(t_s))). \quad (16)$$

Similar to the incremental algorithm, the morphed set of parameters is calculated only at sparsely located excitation points on the "real" time axis. These formulations significantly reduce the computational demand and required memory space.

4.3. Implementation: solution to the third issue

The proposed algorithms consist of transformations that can be calculated before performing actual speech resynthesis. They are $T_{AB}, T_{BA}, T_{Ar}, T_{rA}, T_{Br}$ and T_{rB} and are not dependent on the morphing rates. These transformations and the parameters of the original exemplar materials were calculated at the initialization phase in implementing the proposed algorithms. Since the synthesis procedure of TANDEM-STRAIGHT is already implemented using a piecewise liner interpolation of parameters on the time axis, by replacing the interpolation function with Eqs. 10 (for real-time applications) or 16 (for offline applications), it yields the specialized synthesis procedure for morphing. This implementation significantly reduces memory demand and computational cost.

5. TEMPORALLY VARIABLE MORPHING EXAMPLE

The proposed method was tested using real examples. Two utterances of the Japanese vowel sequence /aiueo/ spoken at fast (voiced duration: 0.44 s) and very slow rates (voiced duration: 1.67 s) were used in the first test. The utterances were spoken by a Japanese male speaker and recorded at a 44.1 kHz sampling rate with 16-bit resolution. Examples with temporally variable morphing rates are presented in a step-by-step manner to illustrate how it works.

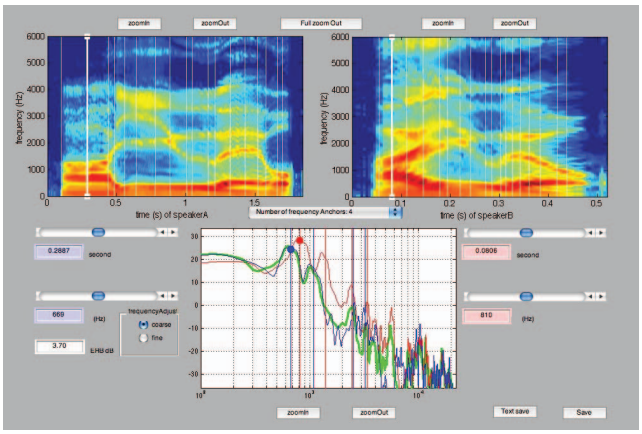

Fig. 1. Interactive editing tool for time-frequency anchoring

5.1. Analysis and anchoring

The TANDEM-STRAIGHT procedure was applied to extract speech parameters of the exemplar utterances. They are calculated every 5 ms (default setting).

An interactive editing tool for anchoring was used to assign anchoring points on the extracted STRAIGHT spectrograms. Figure 1 shows the interface. Two spectrographic representations on top of the interface show the STRAIGHT spectrogram of the slowly spoken utterance and that of the quickly spoken one. Note that the spectrograms only display frequency range up to 6,000 Hz. This is because phonologically important spectral features are located in the displayed frequency range. These spectrograms can be zoomed and dragged by a pointing device (mouse, trackpad etc.)

Vertical lines in the spectrograms represent temporal anchoring points. The thickened line on each spectrogram represents the focus of attention in the time domain. These vertical lines can be dragged by a pointing device or adjusted using the slider attached to each spectrogram. The spectral cross section of each spectrogram at the temporal focus is displayed on the lower plot. The blue line shows the cross section of speaker A (the slowly spoken utterance). The red line shows that of B (the quickly spoken one).

The lines with a large dot represent the focus of attention in the frequency domain. These lines can be dragged by a pointing device or adjusted using the sliders flanking the plot. The frequency axis-morphed and global shape-adjusted version of the cross-sectional spectrum is displayed using the thick green line. The anchoring points are adjusted so as to make the green line fit the blue line in the plot. A dB distance display between this adjusted spectrum and speaker A's spectrum calculated on the ERB_n rate axis (a perceptually derived frequency axis) is presented to help adjustment. The resultant anchoring information was stored and used to prepare temporal axis conversion functions $T_{AB}, T_{BA}, T_{Ar}, T_{rA}, T_{Br}$ and T_{rB} in advance for the actual morphing procedure.

5.2. Definition of temporally variable morphing rates

The first example shows the batch processing algorithm. The temporally variable morphing rate is defined on the reference time axis in this case. The reference time axis is prepared using a temporally constant morphing rate, this time 0.5. The temporally variable morphing rate of the time axis defined on the reference time axis is a straight line starting at 0 and ending at 1.

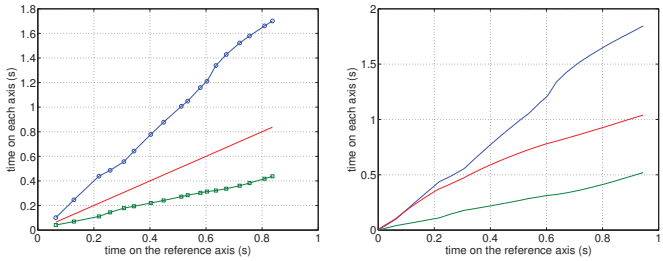

Fig. 2. Reference, original and generated time axes. Blue line: speaker A, dark green line: speaker B, red line of the left plot: reference time and red line in the right plot: "real" time. Circular symbols on lines in the left plot represent temporal anchoring points

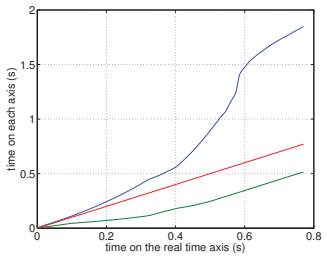

Fig. 3. Time axes generated in the incremental morphing

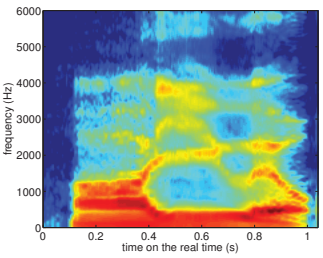

Fig. 4. Temporally morphed STRAIGHT spectrogram

The left hand side of Fig. 2 shows the reference time axis and those of the slowly spoken utterance and the quickly spoken one on the reference time axis. The right hand side of the Fig. 2 shows the time axis of the synthesized speech and the original time axes of the original utterances on the reference time axis.

The blue lines in these plots represent T_{rA}, the mapping from the reference time axis to the slowly spoken time axis. The dark green lines represent T_{rB}, the mapping to the quickly spoken one. The red line in the left plot shows the reference time axis and inevitably is represented as a straight line. The red line in the right plot shows T_{rs}, the mapping to the time axis of the synthesized signal. Note that the initial part of the red line tracks the blue line and the final part of the red line parallels the dark green line. These reflect the desired behavior defined by the temporally variable temporal axis morphing rate.

The second example shows the incremental algorithm. The temporally variable morphing rate changed from 0 to 1 monotonically on the "real" time axis. Note that the length of the synthesized signal cannot be precisely determined in advance. Actually, the morphing rate reached at 1 around the middle of the synthesized utterance and the morphing rate was kept 1 afterwards in this example. Figure 3 shows the temporal axes on the "real" time axis. The blue line in the plot shows T_{sA}, the mapping from the "real" time axis to the slowly spoken time axis. The dark green line shows T_{sB}, the mapping to the quickly spoken one. The red line shows the time axis of the synthesized signal and is inevitably represented as a straight line. Note that again the red line tracks the blue line at the beginning and parallels the dark green line in the end. This is the desired behavior.

5.3. Morphing and resynthesis

Figure 4 shows the STRAIGHT spectrogram using the batch processing algorithm. It illustrates that the proposed method works as intended. The initial part resembles the slowly spoken example and the final part resembles the quickly spoken one. The resynthesized speech signal seamlessly shows transition from the slowly spoken utterance to the quickly spoken one without any artifacts.

Demonstration movies of temporally variable morphing rate manipulations based on the proposed method are linked to our STRAIGHT information page [7]. The page also consists of demonstration movies of TANDEM-STRAIGHT parameter manipulations.

6. DISCUSSION

The manual anchoring procedure is intended for the cases where precise placement of them is crucial, such as psychological stimuli preparation. Automatic procedures for designing conversion functions [8] are under study for musical applications and post-production applications. The proposed method showed graceful behavior when morphing rates were set into the extrapolation region. Perceptual breakdown is also alleviated better by making use of spectral shape decomposition into a global shape and detailed texture [8] and using only the global shape for extrapolation. However, this still needs further study and optimization.

7. CONCLUSION

A generalized framework for auditory morphing with temporally variable morphing rates was proposed. The proposed method also incorporates built-in architecture that alleviates breakdown found in the previous morphing algorithm. The proposed method inherits high quality modification due to STRAIGHT and takes advantage of the computational efficiency of the new TANDEM-STRAIGHT.

8. REFERENCES

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

[2] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27 (3–4), pp. 187–207, 1999.

[3] S. R. Schweinberger, C. Casper, N. Hauthal, J. M. Kaufmann, H. Kawahara, N. Kloth, D. M. C. Robertson, A. P. Simpson, and R. Zeske, "Auditory adaptation in voice perception," *Current Biology*, vol. 18, pp. 684–688, May 2008.

[4] H. Kawahara, T. Ikoma, M. Morise, T. Takahashi, K. Toyoda, and H. Katayose, "Proposal on a morphing-based singing design manipulation interface and its preliminary study," *Journal of Information Processing Society*, vol. 48, no. 12, pp. 3637–3648, 2007, (in Japanese).

[5] T. Yonezawa, N. Suzuki, S. Abe, K. Mase, and K. Kogure, "Perceptual continuity and naturalness of expressive strength in singing voices based on speech morphing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. 9, 2007.

[6] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," in *Proc. ICASSP 2008*. IEEE, 2008, pp. 3933–3936.

[7] H. Kawahara, "STRAIGHT information page," http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/.

[8] M. Onishi, T. Takahashi, T. Irino, and H. Kawahara, "Vowel-based frequency alignment function design and recognition-based time alignment for automatic speech morphing," in *IEEE 2008 Workshop on Spoken Language Technology*. IEEE, 2008, (accepted for publication).