Temporally variable multi-aspect N-way morphing based on interference-free speech representations

Hideki Kawahara*, Masanori Morise†, Hideki Banno‡, and Verena G. Skuk§
*Department of Design Information Sciences, Wakayama University, Wakayama, 640-8510 Japan.
E-mail: kawahara@sys.wakayama-u.ac.jp Tel: +81-73-457-8461
†Faculty of Engineering Computer Science and Engineering, University of Yamanashi, Kofu, 400-8511 Japan.
E-mail: mmorise@yamanashi.ac.jp Tel: +81-55-220-8177
‡Graduate School of Science and Technology, Meijo University, Nagoya, 468-5802 Japan.
E-mail: banno@meijo-u.ac.jp Tel: +81-52-838-2088
§Institute of Psychology, Friedrich Schiller University of Jena, 07743 Jena, Germany.
E-mail: verena.skuk@uni-jena.de Tel: +49-3641-945941

***Abstract*—Voice morphing is a powerful tool for exploratory research and various applications. A temporally variable multi-aspect morphing is extended to enable morphing of arbitrarily many voices in a single step procedure. The proposed method is implemented based on interference-free representations of periodic signals and found to yield highly-naturally sounding manipulated voices which are useful for investigating human perception of voice. The formulation of the proposed method is general enough to be applicable to other representations and easily modified depending on application needs.**

I. Introduction

Voice morphing [1], [2], [3], [4], [5] is a method to generate intermediate voices from two exemplar voices[1]. Voice morphing procedures based on interference-free representations [6], [7] enabled researchers to generate naturally sounding stimulus continuum from representative examples of end points in terms of perceptual or social attributes and yielded many interesting findings [8], [9], [10], [11], [12], [13]. However, its formulation based on binary relations between two examples hindered (or made it very difficult for) applications to general cases such as morphing arbitrarily many voices. This article introduces a novel and general formulation of N-way morphing with descriptions on a Matlab implementation of the algorithm and numerical examples.

II. Background

In this section, former morphing paradigm, temporally variable multi-aspect morphing is briefly reviewed based on its underlying framework TANDEM-STRAIGHT [7]. It is a speech analysis, modification and resynthesis framework consisting of interference-free representations of power spectra [14], [7] and instantaneous frequencies [15], [16], [17], [18] of periodic signals. The new N-way morphing formulation also uses the same representations. Let introduce TANDEM-STRAIGHT and representations first.

This work is partly supported by Grant in Aid for Scientific Research of JSPS and advanced research project of Wakayama University Japan.

[1]The term "*voice* morphing" is used to represent the proposed method and related methods in this article. This types of methods were called speech morphing, sound morphing or auditory morphing in literature.

A. Interference-free representations

TANDEM-STRAIGHT provides a foundation of the proposed method since it allows flexible manipulation of speech parameters without introducing severe degradation of manipulated sounds. The method enables virtually perfect decomposition of speech signals into source related parameters and a filter related spectral representation, called STRAIGHT spectrum. STRAIGHT spectrum does not have any trace of time-frequency periodic variations caused by periodic excitation in voiced sounds. In other words, STRAIGHT spectrum is an "interference-free" representation of power spectrum of periodic sounds [6], [7].

STRAIGHT spectrum is calculated by a two staged procedure. In the first stage, temporal variations caused by periodic excitation in voiced sounds are cancelled out by averaging a pair of power spectra calculated half pitch period apart. This temporally stable power spectral representation is called TANDEM spectrum [14]. In the second stage, F0-adaptive smoothing on the frequency axis is applied to this TANDEM spectrum to suppress periodic variations on the frequency axis. Excessive smoothing caused by applying this anti-aliasing smoother on the already smoothed spectrum (the spectral representation of the time window is the smoother) is compensated based on the consistent sampling theory [19] and adjusted further to improve perceptual quality [20], [21].

The source related information consists of two components. The first one is the fundamental frequency (F0). Several dedicated F0 extractors were developed for older version of STRAIGHT and TANDEM-STRAIGHT [6], [22], [7], [18], [23]. The other one is a set of aperiodicity parameters. Several dedicated representations of aperiodicity were also developed [24], [25], [26]. In the current implementation, they consist of parameters to represent the inflection point and the slope of the sigmoid model, which approximates the spectral shape of random component of the excitation source [27].

B. Temporally variable multi-aspect morphing

Morphing had been defined as interpolation of parameter values between two examples that define end points of the

various morphing trajectories. This definition is straightforward and conceptually simple. However, this simple linear interpolation implementation of morphing fails when temporal axis or frequency axis is extrapolated [5].

The failure is caused by the loss of monotonicity from each example to the morphed time or frequency axis. This loss of monotonicity is prevented by reformulating morphing using log-linear model of derivative of mapping. In this new formulation, derivative of mapping is log-linearly morphed, instead of morphing values directly. Since logarithm of derivative of identity mapping vanishes, morphed time axis $t^{(m)}$ is reduced to the following equation [5].

$$t^{(m)}(t^{(A)}) = \int_0^{t^{(A)}} \left(\frac{dt^{(B)}(\xi)}{d\xi} \right)^{r_{AB}} d\xi, \qquad (1)$$

where $t^{(A)}$ and $t^{(B)}$ represent the time axes of voices A and B respectively. The exponent r_{AB} represents the morphing rate from A to B ($r_{AB} = 0$ yields $t^{(m)} = t^{(A)}$, and $r_{AB} = 1$ yields $t^{(m)} = t^{(B)}$). In this formulation, the time axis of B is a mapping from the time axis of A, $t^{(B)}(t^{(A)})$. An integration variable ξ is a point on the time axis of A.

III. Generalized formulation of morphing

Usually voice morphing is defined as interpolation and/or extrapolation of two exemplar representations [1], [2], [3], [4]. The formulation described in the previous section [5] solved this problem by re-defining morphing using log-linear modeling of derivatives. However, this formulation made further extension of morphing difficult especially when morphing of arbitrary many number of voices is needed.

A. Representation as mapping

The proposed method assumes all representations are not representing simple values. Instead, it assumes that they are representing mapping from an abstract parameter space which is spanned by two coordinates; the abstract time τ and the abstract frequency ν.

Let $\Theta^{(k)}$ represent a set of speech parameters extracted from an indexed utterance, this time k-th voice, for example by using STRAIGHT. Elements of the set $\Theta^{(k)}$ are fundamental frequency (F0) $f_0^{(k)}$, aperiodicity parameter $\boldsymbol{a}^{(k)}$ (a vector consisting of the slope parameter and the inflection frequency of the sigmoidal model) and STRAIGHT spectrogram $P^{(k)}$. In addition to these representations, the coordinate system, time $t^{(k)}$ and frequency $f^{(k)}$, are also the elements of $\Theta^{(k)}$. These elements of the set $\Theta^{(k)}$ are mapping (functions) defined on these abstract coordinates.

$$\Theta^{(k)}(\nu,\tau) = \left\{ f_0^{(k)}\left(t^{(k)}(\tau)\right), \boldsymbol{a}^{(k)}\left(t^{(k)}(\tau)\right), P^{(k)}\left(f^{(k)}(\nu), t^{(k)}(\tau)\right), f^{(k)}(\nu), t^{(k)}(\tau) \right\}. \qquad (2)$$

Figure 1 illustrates this idea using fundamental frequency (F0) as an example. In usual morphing framework, two voices

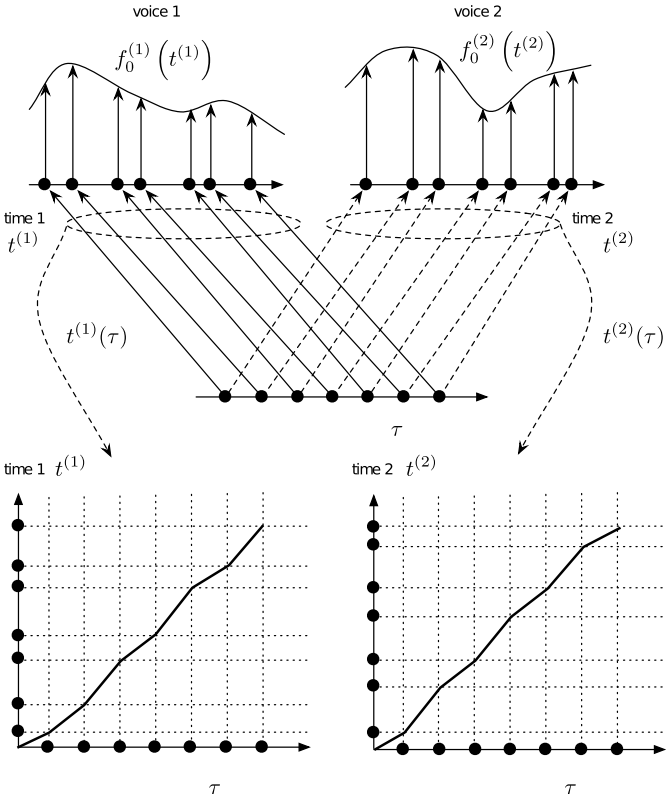

Fig. 1. Schematic diagram to illustrate the idea "representation as mapping" using fundamental frequency (F0) as an example.

are first aligned using the temporal anchors which are represented by black dots on time1 and time 2 in the figure. Then, corresponding values on each aligned point are interpolated (or extrapolated).

In the new formulation, instead of introducing alignment, anchoring points on each time axis are used to define the mapping $t^{(k)}(\tau)$ (k represents the index of the voice) from abstract time τ to time t of each voice as shown in the bottom plots of Fig. 1. Similarly, F0 values of each voice are used to define the mapping $f^{(k)}(t^{(k)})$ from each voice's time t to F0 value f_0.

This is a unified framework. Depending on constraints on mapping, morphing procedures are classified into following three categories.

1) When no constraint is on f_0 value, simple weighted sum (with the constraint on the sum of weights to be one) is relevant to define morphing.
2) When the positivity constraint is required on f_0 value (this is the case for F0 and power spectrum values, for example), log-linear model is relevant to define morphing.
3) When the monotonicity constraint is required on the mapping (for example mapping from the abstract time to time $t(\tau)$ and the mapping from the abstract frequency to frequency $f(\nu)$), log-linear model of the derivative of the mapping is relevant to define morphing.

Failure for fulfilling these constraints yields abnormal morphed sounds or crush of the morphing process, in other words, breakdown. Following sections substantiate algorithms free from these kinds of breakdown.

B. N-way morphing without breakdown

N-way morphing is a procedure to calculate an representation $\Theta^{(m)}(\nu,\tau)$ from a set of representations of voice examples $\Theta^{(1)}(\nu,\tau),\Theta^{(2)}(\nu,\tau),\ldots,\Theta^{(K)}(\nu,\tau)$, where K represents the number of voices. A set of weights W defines contribution of each voice and attribute at each abstract time τ.

$$W=\{\boldsymbol{w}_{F_0}(\tau),\boldsymbol{w}_A(\tau),\boldsymbol{w}_P(\tau),\boldsymbol{w}_{F_x}(\tau),\boldsymbol{w}_{T_x}(\tau)\}, \quad (3)$$

where an element $\boldsymbol{w}_X(\tau)=[w_X^{(1)}(\tau),w_X^{(2)}(\tau),\ldots,w_X^{(K)}(\tau)]^T$ represents a time varying weight vector of a specific attribute $X\in\{F_0,A,P,F_x,T_x\}$. (Suffixes F_0,A,P,F_x and T_x represent F0, aperiodicity parameters, power spectrum (STRAIGHT spectrogram), frequency axis and time axis). To make this definition compatible to conventional morphing, the following normalization constraint has to be satisfied[2].

$$\sum_{k=1}^{K} w_X^{(k)}(\tau)=1. \quad (4)$$

Based on these definitions of symbols, N-way morphing is defined as a transformation T from a set of exemplar voices which are represented as sets of mappings, $\Theta^{(1)}(\nu,\tau),\Theta^{(2)}(\nu,\tau),\ldots,\Theta^{(K)}(\nu,\tau)$, to a specific set of mappings $\Theta^{(m)}(\nu,\tau)$ using a contribution weight W.

$$\Theta^{(m)}(\nu,\tau)=T\left(\Theta^{(1)}(\nu,\tau),\Theta^{(2)}(\nu,\tau),\ldots,\Theta^{(K)}(\nu,\tau);W\right). \quad (5)$$

Time and frequency mapping yielded by morphing always have to be monotonic increasing functions. This condition is fulfilled by defining morphing based on the sum of logarithmic conversion of the derivative of mappings [5].

C. Time and frequency morphing

Since the STRAIGHT parameters are a function of time and frequency, the temporally variable N-way morphing formulations of time and frequency have to be introduced first.

1) Time axis morphing: Let $t^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau))$ represent the morphed time axis with the abstract time τ and $\boldsymbol{w}_{T_x}(\tau)=\left\{w_{T_x}^{(k)}(\tau)\right\}_{k=1}^{K}$ represents the set of (temporally variable) contribution weights on the time axis. The following definition of N-way morphing assures monotonicity of the morphed time $t^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau))$.

$$t^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau))=\int_0^{\tau}\exp\left(\sum_{k=1}^{K} w_{T_x}^{(k)}(\xi)\log\left(\frac{dt^{(k)}(\xi)}{d\xi}\right)\right)d\xi, \quad (6)$$

where $t^{(k)}(\tau)$ represents the time axis of k-th speaker (more precisely, k-th voice) and K represents the number of voices. This equation is reduced to the following form.

$$t^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau))=\int_0^{\tau}\prod_{k=1}^{K}\left(\frac{dt^{(k)}(\xi)}{d\xi}\right)^{w_{T_x}^{(k)}(\xi)}d\xi. \quad (7)$$

Please note that this formulation is the direct extension of (1) to arbitrary many voices. Substituting $K=2, t^{(1)}=t^{(A)}=\xi, t^{(2)}=t^{(B)}$, and $r_{AB}=w_{T_x}^{(2)}$ to (7) yields (1). In other words, the previous formulation [5] is a special case ($K=2$) of this new formulation.

It is not necessary but convenient to make the morphed time axis to have the length of the weighted arithmetic mean of the constituent time axes. Let $t_n^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau))$ represent the standardized morphed time axis.

$$t_n^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau))=\frac{\sum_{k=1}^{K}\left(\int_0^{\tau_M} w_{T_x}^{(k)}(\xi)t^{(k)}(\xi)d\xi\right)}{t^{(m)}(\tau_M;\boldsymbol{w}_{T_x}(\tau_M))}t^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau)), \quad (8)$$

where τ_M represents the larger end of the closed region $[0,\tau_M]$ in the abstract time domain where the voices span.

In real-time applications, this standardization is not necessary. It is because the morphed time axis $t^{(m)}(\tau;\boldsymbol{w}_{T_x}(\tau))$ itself serves as the *real* time axis.

2) Frequency axis morphing: Frequency axis morphing does not consist of time explicitly. In this section, time variable and time-related indices are not explicitly written whenever possible to make formulation simple and easy to read.

Let $f^{(m)}(\nu;\boldsymbol{w}_{F_x})$ represent the morphed frequency axis and $\boldsymbol{w}_{F_x}=\left\{w_{F_x}^{(k)}\right\}_{k=1}^{K}$ represents contribution weight of voices for the frequency axis morphing. Each weight $w_{F_x}^{(k)}$ represents the contribution of the k-th voice[3].

$$f^{(m)}(\nu;\boldsymbol{w}_{F_x})=\int_0^{\nu}\prod_{k=1}^{K}\left(\frac{df^{(k)}(\xi)}{d\xi}\right)^{w_{F_x}^{(k)}}d\xi. \quad (9)$$

For frequency axis morphing, it is necessary to normalize when applying to STRAIGHT based morphing since it is based on DFT and the highest frequency is set to $f_s/2$ where f_s represents the sampling frequency. The normalized morphed frequency $f_n^{(m)}(\nu;\boldsymbol{w}_{F_x})$ is represented as follows.

$$f_n^{(m)}(\nu;\boldsymbol{w}_{F_x})=\frac{f_s}{2f^{(m)}(\nu_M;\boldsymbol{w}_{F_x})}f^{(m)}(\nu;\boldsymbol{w}_{F_x}), \quad (10)$$

where ν_M represents the larger end of the region[4].

It may be more relevant to formulate the frequency axis morphing on the logarithmic frequency. Let $f_N^{(m)}(\nu;\boldsymbol{w}_{F_x})$

[2]This constraint is not necessary for the proposed procedure to function properly. When $\sum_{k=1}^{K} w_X^{(k)}(\tau)>1$, the dynamic range of morphed parameters is expanded. In the opposite case, the dynamic range is compressed.

[3]Note that the weights are constant on the frequency axis. It is possible to generalize for enabling frequency dependent weights. However, it will be too complex to manipulate. Are there any needs?

[4]In usual cases, this corresponds to the Nyquist frequency ($f_s/2$). In our implementation, this anchor is added as a default anchor even when no frequency anchor is assigned.

represent the logarithmic morphing of the frequency axis.[5]

$$f_N^{(m)}(\nu;\boldsymbol{w}_{F_x}) = \exp\left(\int_0^{\nu}\prod_{k=1}^{K}\left(\frac{d\log\left(f^{(k)}(\xi)\right)}{d\xi}\right)^{w_{F_x}^{(k)}}d\xi\right), \quad (11)$$

where $f^{(k)}(0)$ should be positive and $f^{(k)}(\nu)$ should be monotonic increasing function.

The normalized version of the logarithmic morphing $f_{Nn}^{(m)}(\xi;\boldsymbol{w}_{F_x})$ is given below.

$$f_{Nn}^{(m)}(\nu;\boldsymbol{w}_{F_x}) = \frac{f_s}{2f_N^{(m)}(\nu_M;\boldsymbol{w}_{F_x})}f_N^{(m)}(\nu;\boldsymbol{w}_{F_x}), \quad (12)$$

where ν_M represents the larger end of the region.

3) Inverse functions: Morphed time axis and each voice's time axis are represented as functions of the abstract time τ. Similarly, morphed frequency axis and each voice's frequency axis at a certain temporal location are represented as functions of the abstract frequency ν. They are *forward* functions. When synthesizing morphed voice, it is necessary to calculate parameter values on the morphed time and frequency axes. They are calculated firstly mapping coordinates on the morphed time and frequency axes to coordinates on the abstract time and frequency axes using inverse functions of the pre-defined *forward* functions. Then, mapping the abstract coordinates to the coordinates on each voice's time and frequency axes and calculate (by interpolation, usually) desired parameter values. Let substantiate this idea.

Let $\varphi_{T_x}^{(k)}(t^{(m)})$ represent the inverse function from the time on the morphed time axis to the time axis of the k-th voice. It is calculated as a composite function of the inverse function of mapping from abstract time τ to morphed time axis $t^{(m)}$ defined in (6) and the forward function from the abstract time τ to the k-th time axis $t^{(k)}$.

$$\varphi_{T_x}^{(k)}(t^{(m)}) = t^{(k)}(\varphi_\tau(t_{(m)})), \quad (13)$$

where $\varphi_\tau(t_{(m)})$ represents the the inverse function mentioned above. Note that the equation defined by (6) is incrementally calculated. This implies that the inverse function $\varphi_{T_x}^{(k)}(t^{(m)})$ can be calculated also incrementally. This is an important attribute for interactive realtime applications.

Let $\varphi_{F_x}^{(k)}(f^{(m)},t^{(m)})$ represent the inverse function from the frequency on the morphed frequency axis to the frequency axis of the k-th voice at time $t^{(m)}$ on the morphed time axis. Similar to the time axis, it is calculated as a composite function of the inverse function of mapping from abstract frequency ν to morphed frequency axis $f^{(m)}$ defined in (9) and the forward function from the abstract time ν to the k-th time axis $f^{(k)}$.

$$\varphi_{F_x}^{(k)}(t^{(m)}) = f^{(k)}(\varphi_\nu(f_{(m)})), \quad (14)$$

where $\varphi_\nu(f_{(m)})$ represents the inverse function of the forward function $f_{(m)}(\nu)$.

Temporally variable multi-aspect N-way morphing of other parameters, such as 1) fundamental frequency 2) aperiodicity parameters, 3) STRAIGHT spectrogram, are defined using these inverse functions. They are explicitly defined in the following sections.

D. Fundamental frequency (F0) morphing

Let $f_0^{(m)}(t^{(m)};\boldsymbol{w}_{F_0}(t^{(m)}))$ represent the morphed fundamental frequency. The following equation yields the morphed fundamental frequency.

$$f_0^{(m)}(t^{(m)};\boldsymbol{w}_{F_0}(t^{(m)})) = \prod_{k=1}^{K}\left(f_0^{(k)}(\varphi_{T_x}^{(k)}(t^{(m)}))\right)^{w_{F_0}^{(k)}(t^{(m)})}, \quad (15)$$

where $\boldsymbol{w}_{F_0}(t^{(m)})) = \left\{w_{F_0}^{(k)}(t^{(m)})\right\}_{k=1}^{K}$ represents the set of contribution weights for fundamental frequency. Note that the fundamental frequencies are morphed in terms of the logarithmic frequency.

E. Aperiodicity parameter morphing

Let $\boldsymbol{a}^{(m)}(t^{(m)};\boldsymbol{w}_A(t^{(m)}))$ represent the morphed aperiodicity parameter. The following equation yields the morphed aperiodicity parameter.

$$\boldsymbol{a}^{(m)}(t^{(m)};\boldsymbol{w}_A(t^{(m)})) = \sum_{k=1}^{K} w_A^{(k)}(t^{(m)})\boldsymbol{a}^{(k)}(\varphi_{T_x}^{(k)}(t^{(m)})), \quad (16)$$

where $\boldsymbol{w}_A(t^{(m)})) = \left\{w_A^{(k)}(t^{(m)})\right\}_{k=1}^{K}$ represents the set of contribution weights for the aperiodicity parameter.

In addition to this parameter, current aperiodicity representation has additional parameter called "target F0" for designing the sigmoidal shape. It should be morphed logarithmically using the contribution weight $\boldsymbol{w}_A(t^{(m)}))$.

F. STRAIGHT spectrogram morphing

Let $P_{ST}^{(m)}(f^{(m)},t^{(m)})$ represent the generic form of morphed STRAIGHT spectrogram. The inverse function of the frequency mapping $\varphi_{F_x}^{(k)}(f^{(m)})$ represents the inverse function of either the linear frequency morphing $f_n^{(m)}(\nu;\boldsymbol{w}_{F_x})$ or the logarithmic frequency morphing $f_{Nn}^{(m)}(\nu;\boldsymbol{w}_{F_x})$.

$$P_{ST}^{(m)}(f^{(m)},t^{(m)};\boldsymbol{w}_P^{(m)}(t^{(m)})) = \prod_{k=1}^{K}\left(P_{ST}^{(k)}\left(\varphi_{F_x}^{(k)}(f^{(m)},t^{(m)}),\varphi_{T_x}^{(k)}(t^{(m)})\right)\right)^{w_P^{(k)}(t^{(m)})}, \quad (17)$$

where $\boldsymbol{w}_P(t^{(m)})) = \left\{w_P^{(k)}(t^{(m)})\right\}_{k=1}^{K}$ represents the set of contribution weights for the intensity of STRAIGHT spectrogram.

When the STRAIGHT spectrum is represented in terms of dB ($P_{STdB} = 10\log_{10}P_{ST}$) spectrum morphing looks simpler.

$$P_{STdB}^{(m)}(f^{(m)},t^{(m)};\boldsymbol{w}_P^{(m)}(t^{(m)})) = \sum_{k=1}^{K} w_P^{(k)}(t^{(m)})P_{STdB}^{(k)}\left(\varphi_{F_x}^{(k)}(f^{(m)},t^{(m)}),\varphi_T^{(k)}(t^{(m)})\right). \quad (18)$$

[5]The suffix "N" represents John "N"apier, who invented logarithm.

G. Implementation of mapping

The simplest way of defining mapping from the abstract time or frequency to each representation (including axis) is to allocate landmarks on the representation. In this section, an example implementation of the mapping using a piece-wise linear function is presented. The following section introduces an example of time axis mapping.

1) Morphing of the time axis: Let $p^{(k)}(\tau_n)$ represent the temporal location of the n-th anchor point of the k-th speaker. Note that (for making it simple) τ_n is an integer. It is convenient to define $\tau_1 = 0$ and $\tau_{n+1} = \tau_n + 1$, and $p^{(k)}(\tau_1) = 0$ without loosing generality. Then the mapping from the index τ to time t is represented by the following piece-wise linear function $t^{(k)}(\tau)$.

$$t^{(k)}(\tau) = (p^{(k)}(\tau_{n+1}) - p^{(k)}(\tau_n))(\tau - \tau_n) + p^{(k)}(\tau_n). \quad (19)$$
$$\text{for } \ (\tau_{n+1} > \tau \geq \tau_n)$$

Then, its derivative yields as follows.

$$\frac{dt^{(k)}(\tau)}{d\tau} = (p^{(k)}(\tau_{n+1}) - p^{(k)}(\tau_n)), \quad (20)$$
$$\text{for } (\tau_{n+1} > \tau \geq \tau_n) \ .$$

Logarithmic conversion of it is what needed to be used in this extended morphing.

$$\log\left(\frac{dt^{(k)}(\tau)}{d\tau}\right) = \log\left(p^{(k)}(\tau_{n+1}) - p^{(k)}(\tau_n)\right), \quad (21)$$
$$\text{for } (\tau_{n+1} > \tau \geq \tau_n) \ .$$

Then, using the last equation, it is possible to explicitly represent the morphed time axis $t^{(m)}(\tau)$ as follows.

$$t^{(m)}(\tau) = (p^{(m)}(\tau_{n+1}) - p^{(m)}(\tau_n))(\tau - \tau_n) + p^{(m)}(\tau_n),$$
$$\text{for } (\tau_{n+1} > \tau \geq \tau_n) \ , \quad (22)$$

where the coefficient $p^{(m)}(\tau_n)$ is defined by the following.

$$p^{(m)}(\tau_n) = \prod_{k=1}^{K} \left(p^{(k)}(\tau_n) - p^{(k)}(\tau_{n-1})\right)^{\bar{w}_{T_x}^{(k)}(\tau_n)} + p^{(m)}(\tau_{n-1}) \ , \text{ for } \ (n > 1) \ , \quad (23)$$

where

$$\bar{w}_{T_x}^{(k)}(\tau_n) = \frac{w_{T_x}^{(k)}(\tau_n) - w_{T_x}^{(k)}(\tau_{n-1})}{2}, \quad p^{(m)}(\tau_1) = 0$$

Please note that this is only a feasible example of many possible mappings which are usable in the new formulation.

IV. Implementation using Matlab

The new morphing procedures are implemented using Matlab. The inverse functions are implemented using the function `interp1(x,y,xi)` for a one dimensional linear interpolation, where `x` and `y` are coordinate and value for defining the function and `xi` is the location vector to read out the interpolated value.

Figure 2 shows an example implementation of N-way time axis morphing. In this implementation, the function `diff(f)` calculates numerical differentiation of `f` and the function `cumsum(f)` calculates numerical integration of `f`. The first six lines implement (6) directly. In this case, for the sake of simplicity, the contribution weight is constant and normalized.

```
diffTxList = diff(timeAnchorList);
logDiffTxList = log(diffTxList);
targetLength = timeAnchorList(end,:)*wn;
morphedLogDiffTxList = logDiffTxList*wn;
morphedDiffTxList = exp(morphedLogDiffTxList);
morphedTxList = cumsum([0;morphedDiffTxList]);
morphedTxList = ...
      morphedTxList*targetLength/morphedTxList(end);
   . . . . . .
timeMorphedFrame = ...
     interp1(morphedTxList,timeAnchorList(:,ii), ...
     frameOnMorphing);
```

Fig. 2. Example implementation of the time axis morphing. In this case, the contribution weight is temporally constant.

The rest of the lines calculate the inverse function $\varphi^{(k)}(t^{(m)})$. Exchanging a variable `timeAnchorList` representing $t^{(k)}(\tau)$ with a variable `morphedTxList` representing $t^{(m)}(\tau)$ in the argument of `interp1()` and calculating values on the morphed time axis (variable `frameOnMorphing`) yields the desired answer. Elapsed time of a preliminary Matlab implementation of the proposed algorithm is about 0.3 s for three (approximately) 1 s utterances sampled at 44.1 kHz. Elapsed time of the following synthesis stage using morphed parameters was 0.22 s using Matlab (R2012b) on a MacBookPro (2.6 GHz Intel Core i7, 16 GB memory).

V. Numerical demonstrations

An example plots are calculated using a Japanese vowel sequence /aiueo/ spoken by a male speaker in three different expressions (happy, sad and inquiry). Figure 3 shows STRAIGHT spectra of these three expressions. The vertical lines in the plots indicate temporal anchor locations and yellow marks represent frequency anchors. These anchors were manually assigned to align perceptually important spectral landmarks.

In the following sections, demonstration starts from the simplest case, temporally invariant scalar N-way morphing, followed by temporally variable *tied*-aspect N-way morphing. Then, the proposed full version, temporally variable *multi*-aspect N-way morphing is demonstrated.

A. Temporally invariant scalar N-way morphing

In this section, N-way morphing examples are demonstrated with a scalar contributing weight for each attribute and voice. In this case, weights for all voices are represented by a contribution weight vector $\boldsymbol{w} = [w^{(1)}, w^{(2)}, \ldots, w^{(K)}]^T$.

Figure 4 shows the relation between the abstract time τ and each voice $t^{(k)}(\tau)$ using colored lines. It also shows the morphed time axis $t^{(m)}(\tau)$ using a thick black line with a contribution weight vector $[1, 1, 1]^T$ (the normalized version, the variable `wn` in the example list, is $[1/3, 1/3, 1/3]^T$). Please note that the morphed trajectory does not trace averaged value at each abstract time because averaging is on logarithmic converted version of derivative in terms of the abstract time.

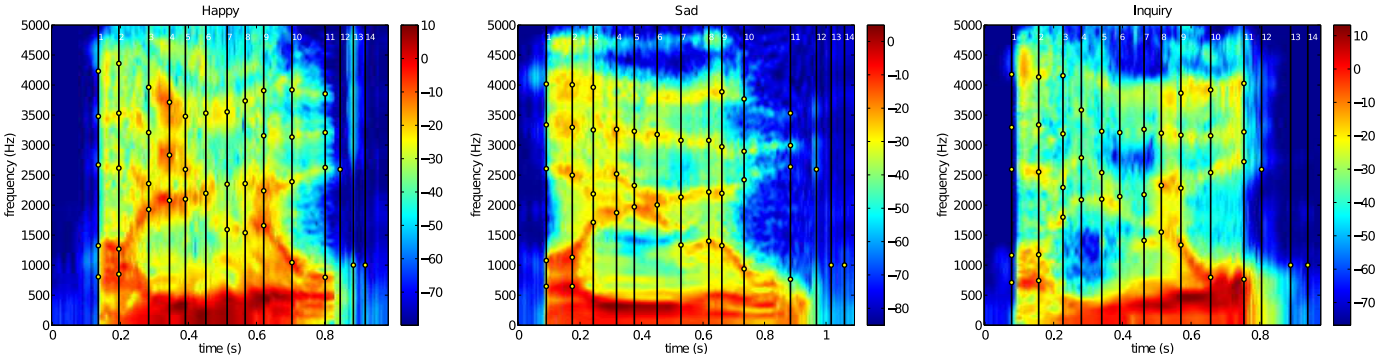

Fig. 3. STRAIGHT spectra of three expressions. From left to right: "Happy," "Sad," and "Inquiry." The vertical lines in the plots indicate temporal anchor locations and yellow marks represent frequency anchors. The spectra shows from 0 Hz to 5 kHz.

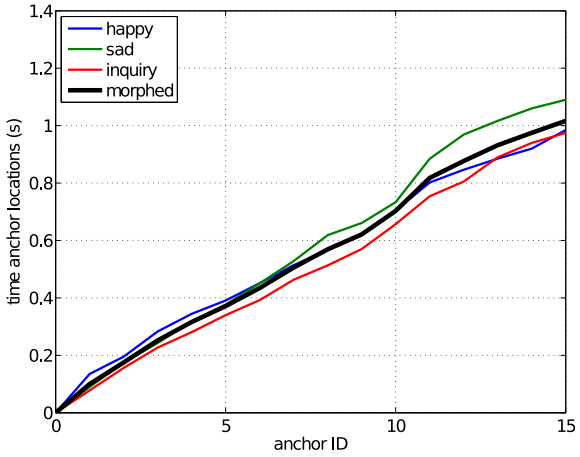

Fig. 4. Mapping from the abstract time τ_n to a speaker's time axes $t^k(\tau)$. The samples are a Japanese vowel sequence /aiueo/ spoken in different expressions (blue: happy, green: sad, red: inquiry). The averaged time axis $t^{(m)}(\tau)$ is also shown in the same plot with a thick black line. The anchor ID$=0$ corresponds the beginning of the voice samples and ID$=15$ corresponds to the end points.

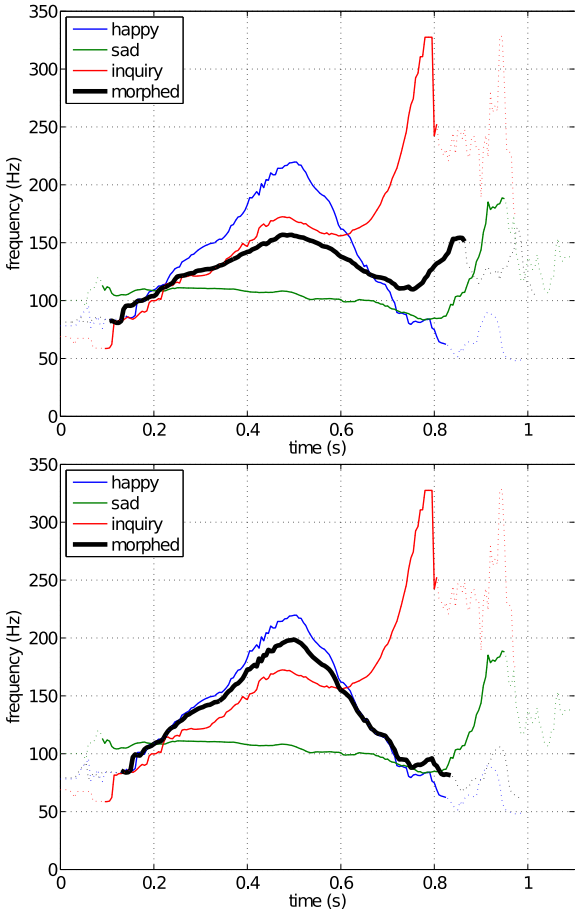

Fig. 5. N-way morphing results of F0 trajectories. Voiced parts are represented using solid lines. The contribution weight vectors are $[1,1,1]^T$ (upper plot) and $[8,1,1]^T$ (lower plot).

Figure 5 shows the original F0 trajectories and the morphed F0 trajectories. The original F0 trajectories are represented using colored lines. The morphed F0 trajectory in each plot is represented by a thick black line. The contribution weight vectors (before normalization) are $[1,1,1]^T$ and $[8,1,1]^T$, from top to bottom. The top plot mixes all expressions equally and keeps the sum of contribution to be unity. This yields *average* voice [11]. The next one is a slightly neutralized version of "happy" expression. Non-unity sum of contribution weights can be used to enhance or suppress weighted values.

B. Temporally variable tied-aspect N-way morphing

The contribution weights for temporally variable multi-aspect N-way morphing is represented as a set of time varying weight vectors for each attribute and defined by (3). Each vector element of an attribute is indexed by the voice index.

Figure 6 shows an example of temporally variable *tied*-aspect N-way morphing, where all attributes have the same time varying contribution weight vector. The contribution weights for each expression is shown as a time series plot in top three plots. F0 trajectory of each expression is represented as a color line in the bottom plot. The morphed F0 trajectory is also shown in the same plot as a thick black line. Four weighting patterns are used. They are "happy-only", "sad-only", "inquiry-only" and "averaged" patterns from left to

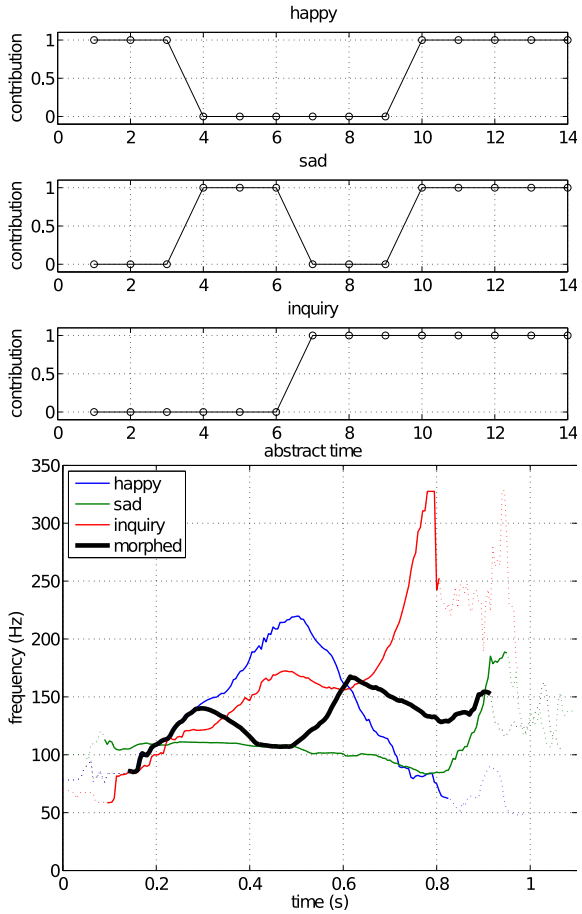

Fig. 6. Temporally variable *tied*-aspect N-way morphing. Upper plot shoes Time series of contribution weights for 'happy', 'sad' and 'inquiry' expressions used in this example. The lower plot shows F0 trajectories for 'happy', 'sad' and 'inquiry' expressions (colored lines) and morphed F0 trajectory (thick black line).

right in the upper plot.

C. Temporally variable multi-aspect N-way morphing

Figure 7 shows examples of temporally variable *multi*-aspect N-way morphing. In this example, the contribution weight for fundamental frequency is the same to the one used in Fig. 6 . The contribution weight for the time axis has non-zero component for only one expression each. They are, from top to bottom, "happy," "sad" and "inquiry" expressions. Note that the morphed trajectory matches the original F0 trajectory where the contribution weight of the expression dominates.

Figure 8 shows the morphed STRAIGHT spectrogram with time and frequency anchor information overlaid. The spectrogram is calculated using the same contribution weights shown in Fig. 6. The spectrogram varies from "Happy," "Sad," "Inquiry" and "Averaged" expressions from left to right. Figure 9 shows the morphed spectrogram with modified contribution weights (from the beginning to the end, "Averaged," "happy," "sad," "inquiry" and "Averaged" expressions).

The sound examples including additional materials are linked to our demonstration web page [28].

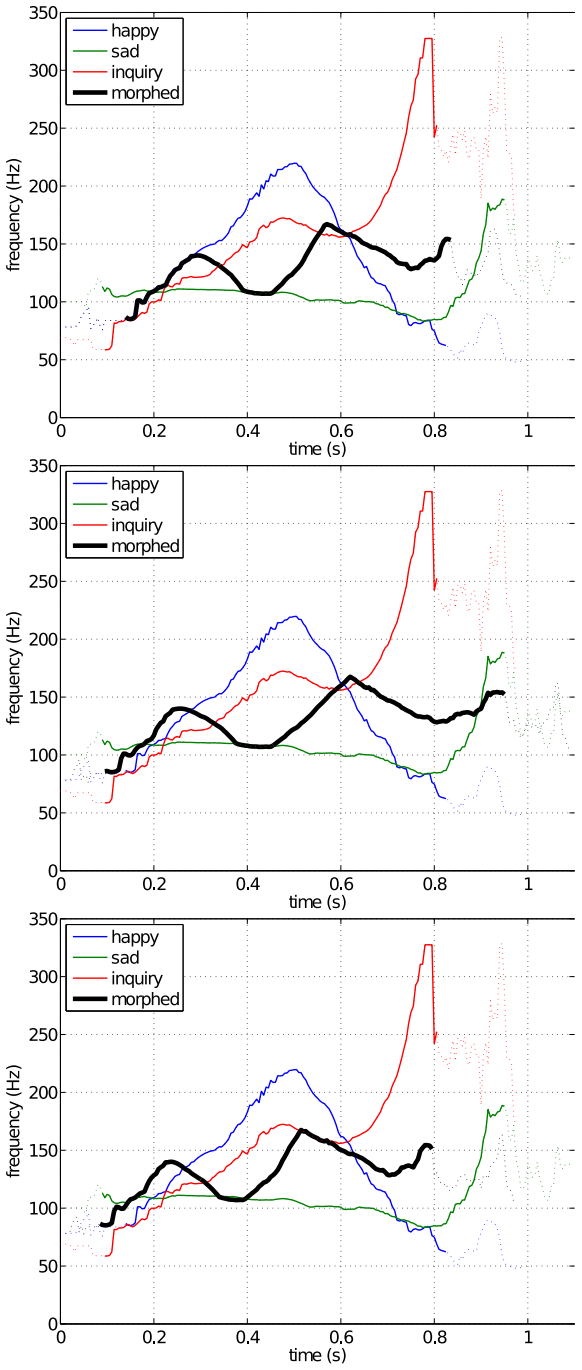

Fig. 7. Temporally variable *multi*-aspect N-way morphing. The contribution weight for fundamental frequency trajectories is the same to Fig. 6. The contributing weight for the time axis of each plot has non-zero values only for 'happy', 'sad' and 'inquiry' time axes respectively.

1) Application and future directions of this flexible framework: This single shot procedure of N-way morphing minimizes possible quality degradation of morphed speech. This is desirable for making averaged voices [11] and/or caricature voices.

In addition to this basic form of application, an extension for making this procedure more flexible is introduced. The

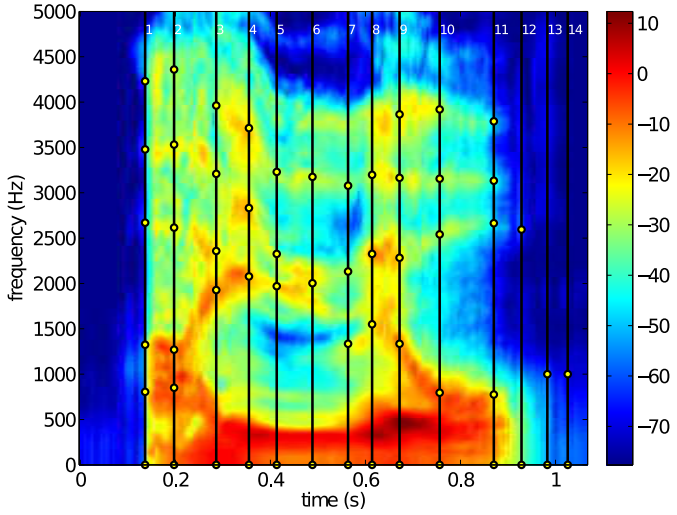

Fig. 8. Morphed STRAIGHT spectrogram with morphed anchors. The contribution weights are the same to Fig. 6. (From the beginning to the end, "happy," "sad," "inquiry" and "Averaged" expressions.)

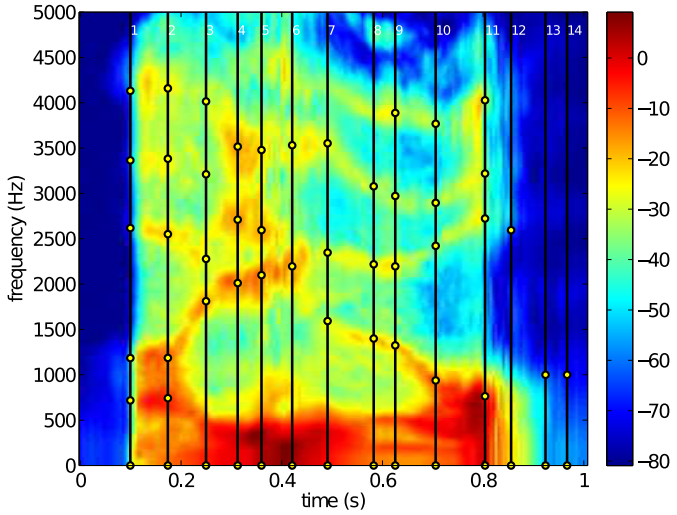

Fig. 9. Morphed STRAIGHT spectrogram with morphed anchors. The contribution weights are designed to yield, from the beginning to the end, "Averaged," "happy," "sad," "inquiry" and "Averaged" expressions.

output of the morphing procedure was formatted to have the compatible structure with STRAIGHT objects. It means that the morphed result can be re-used as one of inputs to the morphing procedure. In general use, this cascading use of morphing procedure is not very practical in terms of sound quality, since morphing procedure inevitably introduce slight degradation and cascading use accumulates degradation in each stage. However, this unified representation makes manipulation of speech parameters flexible. By introducing dummy STRAIGHT objects, and removing constraint on the sum of contribution weight, this temporally variable multi-aspect N-way morphing procedure provides a unified means to manipulate STRAIGHT parameters. For example, modification of each formant frequency and/or bandwidth independently [29]

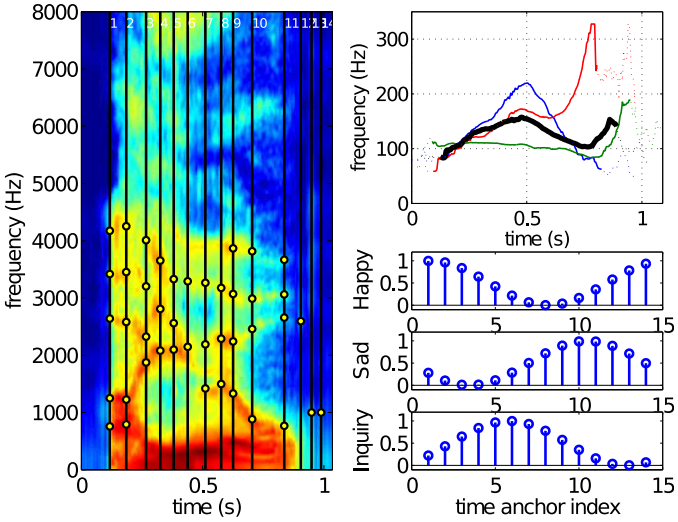

Fig. 10. Snapshot of one of demonstration movies. The animation movies are linked to the demonstration web page [28].

can be implemented very easily.

In this implementation, mapping from abstract time and frequency (index) space to the each voice's time and frequency coordinates is defined by manually assigning anchoring points on each STRAIGHT spectrogram. This procedure is very time consuming and requires fundament understanding of speech production mechanism, speech perception, acoustic phonetics and digital signal processing. This demanding procedure was intentionally introduced [3] because using this tool for exploratory research, researchers have to have full control of all contributing parameters and awareness of possible artifacts.

However, for wider applications such as post processing of audio part of multi-media contents, this heavy reliance on manual procedure is prohibiting. Forced alignment of temporal anchors using speech recognition engines [30], [31] is a promising approach. Frequency axis normalization based on interference-free power spectral representation [32] is also effective to reduce assignment cost. In addition to these semi-automatic procedures, fully automatic morphing procedure is a valuable and important goal for further research.

Another important topic is manipulator design. Figure 11 shows a prototype GUI for temporally static tied aspect three-way morphing. A slightly transparent gray sphere is used as the manipulation knob to control a set of three dimensional morphing rates. Area of triangles formed by the knob location and three vertices are used as the morphing ratios. By using a three dimensional pointing device, up to four-way morphing, unique mapping from physical coordinate of one knob to morphing rates is possible. But for $N > 4$ voices morphing, it is impossible to design an isomorphic continuous mapping scheme in a three dimensional physical space. In such many voice cases, dedicated design for each specific application has to be introduced. These are interesting topic for further research.

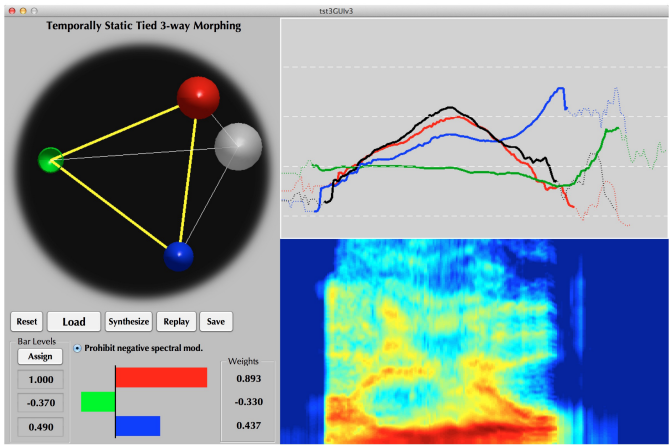

Fig. 11. Morphing rate manipulation GUI for three-way temporally static tied-aspect morphing.

VI. Conclusions

A simple and flexible framework for enabling morphing arbitrarily many voices is proposed. It allows to assign time varying contribution weights to all voices and all attributes independently. This formulation is applicable to realtime handling of contribution weights as well as interactive-offline handling such as post production process of multi-media. In spite of conceptual simplicity of the proposed method, it is a time consuming, hard and prone-to-error task for allocating anchoring points for each analysis results. Semi-automatic interactive GUI is crucially important for making the proposed method accessible.

Acknowledgment

The authors thank Pascal Belin for motivating us for developing direct N-way morphing for the first time. They also thank Stefan Schweinberger for providing us to have chance to concentrate on this topic again and encouraging us by suggesting exciting application possibilities. The authors also thank Minoru Tsuzaki and Tomoyasu Nakano for comments and discussions.

References

[1] M. Slaney, M. Covell, and B. Lassiter, "Automatic audio morphing," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, vol. 2, pp. 1001–1004.

[2] M. Abe, "Speech morphing by gradually changing spectrum parameter and fundamental frequency," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, vol. 4, pp. 2235–2238 vol.4.

[3] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Proc. ICASSP2003*, Hong Kong, 2003, vol. I, pp. 256–259.

[4] Hui Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1301–1312, 2006.

[5] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," in *Proc. ICASSP2009*. IEEE, 2009, pp. 3905–3908.

[6] 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, no. 3-4, pp. 187–207, 1999.

[7] 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. ICASSP2008*, 2008, pp. 3933–3936.

[8] David R. R. Smith, Thomas C. Walters, and Roy D. Patterson, "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3628–3639, 2007.

[9] Peter F. Assmann and Terrance M. Nearey, "Relationship between fundamental and formant frequencies in voice preference," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. EL35–EL43, 2007.

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

[11] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G.A. Rousselet, H. Kawahara, and P. Belin, "Vocal attractiveness increases by averaging," *Current Biology*, vol. 20, no. 2, pp. 116–120, 2010.

[12] R. Zäske S. R. Schweinberger and H. Kawahara, "Voice aftereffects of adaptation to speaker identity," *Hearing Research*, vol. 268, pp. 38–45, 2010.

[13] V. Skuk and S. R. Schweinberger, "Influences of fundamental frequency, formant frequencies, aperiodicity and spectral level information on the perception of voice gender," *Journal of Speech, Language, and Hearing Research*, 2013, (In press).

[14] M. Morise, T. Takahashi, H. Kawahara, and T. Irino, "Power spectrum estimation method for periodic signals virtually irrespective to time window position," *Trans. IEICE*, vol. J90-D, no. 12, pp. 3265–3267, 2007, [in Japanese].

[15] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, pp. 1493–1509, November 1966.

[16] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520 – 538, 1992.

[17] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. II. algorithms and applications," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 540 – 568, 1992.

[18] H. Kawahara, T. Irino, and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," *ICASSP2011*, pp. 5420 –5423, may 2011.

[19] Michael Unser, "Sampling–50 years after Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, 2000.

[20] H. Akagiri, M. Morise, T. Irino, and H. Kawahara, "Evaluation and optimization of F0-adaptive spectral envelope estimation based on spectral smoothing with peak emphasis," *Trans. IEICE*, vol. J94-A, no. 8, pp. 557–567, 2011, [in Japanese].

[21] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *SADHANA*, vol. 36, no. 5, pp. 713–722, 2011.

[22] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free f0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Interspeech'05*, Lisboa, 2005, pp. 537–540.

[23] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulatio," in *Interspeech2012*, 2012, Session: O2d.05.

[24] Hideki Kawahara, "Exemplar-based voice quality analysis and control using a high quality auditory morphing procedure based on straight," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[25] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, J. C. Williams, and M. Morise, "Noh voice quality," *J. Logopedics Phoniatrics Vocology*, vol. 34, no. 4, pp. 157–170, 2009.

[26] Hideki Kawahara, Masanori Morise, Toru Takahashi, Hideki Banno, Ryuichi Nisimura, and Toshio Irino, "Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems," in *Proc. Interspeech2010*. ISCA, 2010, (in print).

[27] H. Kawahara and M. Morise, "Simplified aperiodicity representation

for high-quality speech manipulation systems," *Proc. ICSP 2012*, vol. 1, pp. 579–584, 2012.

[28] "http://www.sys.wakayama-u.ac.jp/%7ekawahara/APSIPA2013/," 2013.

[29] Chang Liu and Diane Kewley-Port, "Vowel formant discrimination for high-fidelity speech," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 1224–1233, 2004.

[30] "The Hidden Markov Model Toolkit (HTK) ," http://htk.eng.cam.ac.uk.

[31] "Open-source large vocabulary CSR engine Julius," http://julius.sourceforge.jp/en_index.php?q=index-en.html.

[32] M. Kobayashi, R. Nisimura, T. Irino, and H. Kawahara, "Estimated relative vocal tract lengths from vowel spectra based on fundamental frequency adaptive analyses and their relations to relevant physical data of speakers," in *Proc. ICA 2013*, Montreal, 2013, p. 5aSCb44.