SPEECH MORPHING BASED ON BIOLOGICALLY RELEVANT SIGNAL REPRESENTATIONS

Hideki Kawahara
Auditory Media Laboratory, Faculty of Systems Engineering, Wakayama University, Wakayama, Japan

Abstract: Voice morphing based on a high fidelity VOCODER is a unique strategy to explore attributes which are closely related to biological states of speakers. The method is based on a temporally stable power spectral representation and spectral envelope recovery based on a new formulation of the sampling theory. The morphing algorithm itself is re-formulated to enable extrapolation without introducing perceptual and objective breakdown. It also extended to make temporally-variable multi-aspect morphing possible. GUI (graphical user interface) based tools are implemented to handle complexities introduced by these extensions. For characterizing voicing, a bottom-up procedure to scoop all local repetition detector, a residual-based irregularity detector and a group delay-based acoustic event detector with multi-resolution analysis are prepared.

Keywords:— Spectrum, periodicity, speech perception, voicing, morphing

I. INTRODUCTION

Repetitive structures [1] play important roles in biological systems from animal calls to voiced sounds in human speech. However, usual short term Fourier based analysis methods including cepstrum and LPC analyses, suffer from interferences caused by this repetitive structure. Recently, a simple method for calculating interference-free power spectra [2] and new formulation of sampling theory [3] led to an invention of a speech analysis, modification and synthesis procedure called TANDEM-STRAIGHT [4]. TANDEM-STRAIGHT consists of a new bottom-up procedure to scoop all local repetitive structure based on this power spectral representation. TANDEM-STRAIGHTe was also applied to extend speech morphing procedure [5] and yielded a temporally variable multi-aspect morphing procedure [6]. These new set of procedures are integrated with visualization and GUI tools [7] to provide a strong basis for investigating biomedical aspects of voice emission and perception.

II. POWER SPECTRUM OF PERIODIC SIGNALS

Assume that a repetitive signal \(x(t)\) has a fundamental period \(T_0\) and its short term Fourier transform \(S(\omega, t)\) is calculated using a time window \(w(t)\). With a mild conditions on the window function \(w(t)\), the following power spectrum \(P_T(\omega, t)\) does not have temporal variations due to periodicity (repetition).

\[
P_T(\omega, t) = |S(\omega, t - T_0/4)|^2 + |S(\omega, t + T_0/4)|^2. \quad (1)
\]

\(P_T(\omega, t)\) is called TANDEM spectrum afterwards. This operation does not have impact on frequency resolution of the original time windowing. Typical selection of the time window is a Blackman window having its duration set 2.5\(T_0\).

A. Periodic variations in the frequency domain

Variations of \(\log(P_T(\omega, t))\) in the frequency domain is closely approximated by an additive sinusoid with a period \(\omega_0 = 2\pi f_0\), where \(f_0\) represents fundamental frequency (F0). It is completely eliminated applying a frequency domain smoother that has zeros at \(n\omega_0\) on its spatial frequency transfer function. The simplest one is a rectangular smoother with \(\omega_0\) for its width. Let this smoother \(h_1(\omega; \omega_0)\). Reasonable windows convolved with this smoother also have zeros at the same place and can be used for the same purpose. Let’s call this F0 adaptively smoothes power spectrum “smoothed spectrum” \(P_S(\omega, t)\).

B. Consistent sampling

Unfortunately, \(P_S(\omega, t)\) does not precisely agree with Fourier transform of the (hypothetical) unit waveform that is repeated. There are two sources of smearing for the unit waveform. One is frequency response of the time window and the other is the smoothing function. Consistent sampling provides a way to solve this problem. A correction digital filter \(Q(z)\) in the frequency domain can be designed using convolution of \(h_1(z)\) and \(W(z)\), where they are \(h_1(\omega; \omega_0)\) and \(w(t)\) represented in terms of \(z\) transform.

\[
Q(z) = \frac{1}{a(z)}. \quad (2)
\]

Partially supported by Grants-in-Aid for Scientific Research (A)19200017 by JSPS and CrestMuse project of JST.
where $a(z)$ is the convolution of $h_1(z)$ and $W(z)$. Please note that the polynomial $Q(z)$ has infinite number of coefficients, because $a(z)$ is effectively a function with a finite support. Approximation errors between harmonic frequencies is dependent on the effective interpolating function that is convolution of $h_1(z)$ and $W(z)$, this case. In other words, there is a room for improvement in designing the smoothing function, actually a triangular function that is convolution of $h_1$ and $h_1$ is a better smoother for speech sounds.

C. STRAIGHT spectrum

Taking into account of the fact that $\log(1 + x) \approx x$ when $|x| \ll 1$ and absolute value of coefficients of $Q(z)$ decreases very rapidly, smoothed spectrum that preserves values at harmonic frequencies $P_{TST}(\omega, t)$ is calculated using the following equation.

$$P_{TST}(\omega) = e^{(\tilde{q}_1(L(\omega - \omega_0) + L(\omega + \omega_0))) + \tilde{q}_0 L(\omega))}, \quad (3)$$

where $L(\omega) \equiv \log(P_S(\omega))$ and $\tilde{q}_0, \tilde{q}_1$ are truncated and adjusted version of the coefficients of $Q(z)$. Please note that variable $t$ is not represented in this equation to make appearance simple. Afterwards, whenever not confusing, the same practice applies. This spectrum does not have trace of periodicity while it preserves spectral values at harmonic frequencies. This spectrum is called STRAIGHT spectrum, because it is virtually identical to the spectrum calculated using the legacy-STRATIST [1].

III. LOCAL PERIODICITY DETECTOR

Since $P_S(\omega)$ is the periodicity eliminated version of $P_T(\omega)$ and the effect of periodicity is multiplicative, dividing $P_T(\omega)$ by $P_S(\omega)$ leaves a constant $c_0$ and the periodic component $P_C(\omega)$.

$$P_C(\omega) = \frac{P_T(\omega)}{P_S(\omega)} - c_0. \quad (4)$$

Ideally, Fourier transform of $P_C(\omega)$ has a unique peak at $\tau = T_0$. However, in practice, low S/N in lower frequency region and FM side-bands in F0 varying speech (that is usually the case) direct application of Fourier transform on $P_C(\omega)$ yields erroneous and noisy results.

To investigate vocal fold vibration, it is better to select the base-band frequency region using a frequency domain weighting function. The frequency weighting function can be located anywhere depending on aspects to be investigate, for example, to investigate regularity of glottal closure instant, the function can be centered around 3 kHz. The weighting function for selecting base-band region $w_{\omega_0,N}(\omega)$ has the following form and defined in $[-N\omega_0, N\omega_0]$.

$$w_{\omega_0,N}(\omega) = c_1 \left(1 + \cos\left(\frac{\pi \omega}{N\omega_0}\right)\right), \quad (5)$$

where $c_1$ is a normalization constant. Then, Fourier transform of the windowed version of the periodic component has a less noisy peak at $\tau = T_0$.

$$A(\tau; T_0) = \int_{-\infty}^{\infty} w_{\omega_0,N}(\omega) P_C(\omega) e^{-j\omega \tau} d\omega, \quad (6)$$

where the assumed period is explicitly denoted in $A(\tau; T_0)$. Increasing $N$ sharpens the peak and makes it tolerant to background noise while makes it susceptible to FM and AM meaning that there is a trade-off relation.

The designed detector is specialized to the assumed $T_0$. By assuming periods $T_{0k}, (k = 1, \ldots, M)$ systematically on the logarithmic lag axis, they are combined to cover periodicity range of interest using the following equations.

$$\tilde{A}(\tau) = c_2 \sum_{k=0}^{M} w_{LAG}(\tau; T_{0k}) A(\tau; T_{0k}), \quad (7)$$

$$w_{LAG}(\tau; T_{0k}) = 1 + \cos\left(\frac{\pi \log_2(\beta\tau/T_{0k})}{\beta\tau/T_{0k}}\right), \quad (8)$$

where a constant $c_2$ is adjusted for $\tilde{A}(\tau)$ to have a value 1 for periodic signals. The weighting function $w_{LAG}(\tau; T_{0k})$ defined in $[-T_{0k} < \beta\tau < T_{0k}]$ is use to suppress spurious peaks in $A(\tau; T_{0k})$ by adjusting selectivity using $\beta$. Typical selection of $T_{0k}$ has the following form.

$$T_{0k} = T_L 2^{-k/N_c} \quad (9)$$

where $N_c$ determines the number of specialized detectors in one octave and $T_L$ represents the longest period to be investigated.

Peaks of periodicity measure $\tilde{A}(\tau)$ represent local repetitions of waveform using the best time-frequency resolution in each period scale. It is a bottom-up exhaustive periodicity detection system to be used to characterize repetitive structures in speech. Therefore, this method is called XSX (eXcitation Structure eXtractor) [8].

IV. APERIODICITY REPRESENTATION

The XSX is able to extract several types of aperiodicity such as jitter and shimmer as timing fluctuations and amplitude fluctuations of each excitation event respectively. However, there still remains other types of deviations from precise repetition. Linear prediction residuals from around a repetition period apart (both forward and backward), calculated on two types of time axes, are
used to represent aperiodic component that cannot be represented by XSX. Two types of time axes are as follows. The first one is the usual time axis. The second one is a warped time axis that is stretched in proportion to its instantaneous frequency corresponding to the repetition period. This selection of the second time axis makes apparent repetition period constant. The smaller residual of these two predictions is used as an index to represent aperiodicity in each time-frequency band region. Octave division of frequency band is used in the current implementation with keeping the narrowest bandwidth wider than 500 Hz.

For diagnostic applications, an acoustic event and group delay based representation [9] is also used. However, it still is one of the future topics to integrate this event based representation into TANDEM-STRAGHT and morphing system.

V. TIME-VARIABLE MULTI-ASPECT MORPHING

Speech morphing was originally designed [5] based on linear interpolation and extrapolation of parameter values. This definition was found fragile when parameters are extrapolated. A new definition that enables time-variable multi-aspect morphing was proposed by re-defining morphing, based on linear interpolation in the logarithm of derivative domain [6].

Using this formulation, let $T_{Am}(x_A)$ represent a morphing transformation of a parameter $x_A$ of example A to parameter $x_m$ on the morphing axis $m$. A temporally variable morphing rate for the parameter $r_{AB}(t)$ is defined to have the value 0 when the morphed result is equivalent to example A and to have the value 1 when the morphed result is equivalent to example B. A new morphing definition is introduced and described using this notation.

To alleviate breakdown in explorative morphing, morphing is redefined based on a logarithm of the derivative of mapping functions. This new definition of morphing also makes the morphing procedure simpler as follows:

$$ T_{Am}(x_A) = \int_0^{x_A} \exp \left( \log \left( \frac{dT_{Am}(\lambda)}{d\lambda} \right) \right) d\lambda $$

$$ = \int_0^{x_A} \exp \left( (1 - r_{AB}(\lambda)) \log \left( \frac{dT_{AA}(\lambda)}{d\lambda} \right) \right) d\lambda $$

$$ + r_{AB}(\lambda) \log \left( \frac{dT_{AB}(\lambda)}{d\lambda} \right) d\lambda $$

$$ = \int_0^{x_A} \left( \frac{dT_{AB}(\lambda)}{d\lambda} \right) r_{AB}(\lambda) d\lambda, \quad (10) $$

because logarithmic conversion of the identity mapping vanishes. This formulation assures monotonicity of $T_{Am}$ if the coordinate conversion $T_{AB}$ from speaker A to B is monotonic.

Two morphing algorithms are formulated based on this new definition of morphing: real-time morphing and off-line morphing. In the case of real-time morphing, the morphing rates are incrementally supplied and used to update morphed parameters incrementally. This formulation is useful for interactive applications.

In the case of off-line morphing, the morphed time axis, which is also the time axis for the morphed signal, is calculated for the first time. Then, other morphed parameters are calculated using the morphing rate on this new reference axis. This formulation is necessary for psychophysical stimuli preparation and biomedical diagnostic applications. Fig. 1 illustrates the synthesis procedure using these parameters and transformations.

VI. GUI TOOLS

GUI tools are equipped with analysis tools and visualization interface. Fig. 2 shows GUI for F0 extraction. The default F0 extractor is XSX. In other words, it is not a mere F0 extractor. It is a visualization tool for excitation
Fig. 2. GUI tool for XSX analysis. Top panel shows periodicity index for each repetition structure. The middle panel shows extracted local repetition structure in terms of frequency. The bottom panel shows the waveform. The sample is /haï/ (Yes, in English) spoken with strong anger. (In originally figure, marks are color coded in order of repetition salience.)

structure analyses.  

The middle plot of the display shows localized repetitions in terms of frequency. This plot can be zoomed both in time and frequency and can be dragged to center interesting regions. The thick gray line (cyan in color) represents the most salient periodicity that corresponds to F0. Application of XSX to Noh voice analyses illustrated interesting subharmonic structure and non-classical transition of the most salient periodicity [8].

VII. DISCUSSION

The GUI tools and underlying algorithms is designed to promote exploratory research strategy for investigating phenomena they are difficult to be categorized a-priori. By using morphing to generate a set of stimulus continuum combined with after effect, auditory adaptation in voice perception was discovered [10]. In other words, a stimulus continuum generated using morphing provides means to objectively quantify non- or pre-categorical percept and phenomena. One prospective example is evaluation of healing process from vocal fold surgery.

VIII. CONCLUSION

A temporally-variable multi-aspect morphing method based on a temporally stable representation of periodic signals combined with a bottom-up repetitive structure extractor and a residual-based aperiodicity extractor are introduced. This algorithm and a set of dedicated GUI tools provide a strong basis for exploratory research on biomedical aspects of voice emission and perception.

REFERENCES


