ANALYSIS AND SYNTHESIS OF STRONG VOCAL EXPRESSIONS: EXTENSION AND APPLICATION OF AUDIO TEXTURE FEATURES TO SINGING VOICE

Hideki Kawahara[*]

Faculty of Systems Engineering
Wakayama University
Wakayama, Wakayama, 640-8510 Japan

Masanori Morise[†]

College of Information Science and Engineering
Ritsumeikan University
Kusatsu, Shiga, 525-8577 Japan

ABSTRACT

Realistic reconstruction and manipulation of strong vocal expressions found in singing voices is a challenging and exciting topic. A speech analysis, modification and resynthesis framework based on interference-free power spectral and instantaneous frequency representations for periodic sounds is extended for handling such voices. Strong expressions are typically characterized by rapid variations in excitation timing and strength as well as complex structured excitation. Three types of excitation source extractors are revised and introduced to handle them. Preliminary tests successfully replicated strong vocal expressions. Also, additional attribute representations for modifying excitation and spectral information based on audio texture features are briefly discussed.

Index Terms— speech analysis, speech synthesis, Human voice, Vocoders, periodic structures

1. INTRODUCTION

Singing voice is a unique musical instrument not only because it can convey lyrics and melody at the same time, but also because voice itself is extremely expressive and is coupled with basic instinct and biological status. Singers exploit full range of these advantages in their performance. Their exploitation sometimes results into unique strong vocal expressions associated with irregularities and complex excitation structures.

This article introduces a set of textural representations for characterizing these expressive attributes and proposes a framework for flexible manipulation of expressive attributes. The proposed framework is implemented as an extension to TANDEM-STRAIGHT [1], a speech analysis, modification and resynthesis system.

The following sections start from brief introductions to concept and implementation of TANDEM-STRAIGHT and temporally variable multi-aspect morphing [2]. Then, discussions on the extension for strong vocal expressions are presented with numerical examples.

2. BACKGROUND: TANDEM-STRAIGHT

In TANDEM-STRAIGHT and its predecessor STRAIGHT (legacy-STRAIGHT [3]), quasi periodic excitation of voiced sounds is understood as a mechanism for sampling underlying smooth time-frequency representations which cannot be observed directly. This repetitive excitation is effective for making voices stand out loud and salient in pitch. However, this repetitive excitation is troublesome for usual short term Fourier analysis (STFT), because it introduces periodic variations both in the time and the frequency domains. TANDEM-STRAIGHT removes these variations in two steps: temporally stable power spectrum representation and spectral envelope recovery based on a new sampling theory.

[*]Partly supported by Grants-in-Aid for Scientific Research 22650042 by JSPS and Advanced Research Initiative by Wakayama University.

[†]Partly supported by Ono Acoustics Research Fund and Grants-in-Aid for Scientific Research 23700221 by JSPS.

2.1. Temporally stable power spectrum

Averaging power spectra calculated at two temporal locations which are one half-pitch period apart eliminates temporal variations due to periodicity [4, 1]. This procedure is applicable to wide range of windowing functions with reasonably low side-lobe levels and with an equivalent pass band spanning up to two harmonic components. In the current implementation of TANDEM-STRAIGHT, a Blackman window having an F0-adaptive window length of $2.5T_0$ is used, where T_0 represents the fundamental period. Detailed discussions of windowing functions for this procedure are given in the literature [5].

2.2. Spectral envelope recovery

Periodic excitation in the time domain is periodic sampling in the frequency domain. This interpretation and consistent sampling theory [6] provide a procedure to recover spectral envelope which does not have periodic variations due to frequency sampling and preserves power spectral levels at harmonic frequencies [1]. Consistent sampling theory is crucially important in spectral envelope recovery, because vocal tract transfer functions are not band-limited.

Recently, a new implementation of this procedure based on logarithmic conversion of spectral information was introduced to improve recovery accuracy and perceptual sound quality [5, 7]. In this implementation, spectral envelope $P_{ST}(\omega)$ is calculated from the temporally stable power spectrum $P_T(\omega)$ using the following equation:

$$P_{ST}(\omega) = \exp\left(\mathcal{F}\left[g_1(q)g_2(q)C_T(q)\right]\right), \quad (1)$$

where $C_T(q)$ represents cepstrum of $P_T(\omega)$ and q represents quefrency. Symbol $\mathcal{F}[\,]$ represents Fourier transform. Two lifters $g_1(q)$ and $g_2(q)$ are defined below:

$$g_1(q) = \tilde{\alpha}_0 + 2\tilde{\alpha}_1 \cos\left(2\pi q f_0\right), \quad (2)$$

$$g_2(q) = \frac{\sin(\pi f_0 q)}{\pi f_0 q}, \quad (3)$$

where $f_0 = 1/T_0$ represents fundamental frequency (F0). The second lifter $g_2(q)$ represents F0-adaptive spectral smoothing using rectangular smoothing function (width is set to f_0). The first lifter $g_1(q)$ represents a digital filter on the frequency axis for compensating over-smoothing due to $g_2(q)$ and time windowing used to calculate $P_T(\omega)$. Detailed discussion on this cepstrum-based implementation is given in the literature [5].

The filter coefficients $\tilde{\alpha}_0, \tilde{\alpha}_1$ of $g_1(q)$ are numerically determined to minimize the Itakura-Saito spectral distance on the perceptual frequency axis (ERB_N number [8]) for various types of vocal tract shapes [9] and excitation source variations using LF-model [10]. Subjective tests revealed that this new implementation using the optimized coefficients $\tilde{\alpha}_0 = 1.18, \tilde{\alpha}_1 = -0.09$ provides better manipulated speech sounds [7] than legacy-STRAIGHT and the first implementation of TANDEM-STRAIGHT [1] as well as PSOLA[11]. Since the final form of this implementation is Cepstrum liftering, it is interesting to investigate on relations with latest True-Envelope based approach [12].

2.3. Excitation source representations

To resynthesize speech sounds, the extracted (underlying smooth) time-frequency representation has to be excited using source related information. This is the weakest part of STRAIGHT and extensions for strong expressions are mainly dealing with this representation.

STRAIGHT intentionally discards the original waveform information (in other words phase information) in the analysis stage. This is because one of the important goal in designing STRAIGHT is to use perceptually relevant information only. Waveform preservation is not necessary for perceptually identical speech reproduction, because two independent segments of Gaussian white noise having the same variance are perceptually indistinguishable.

This phase negligence and temporally stable power spectral representations eliminate need of "pitch marking," which is prerequisite for conventional pitch synchronous methods and is fragile. In other words, TANDEM-STRAIGHT is a "pitch-marking-free pitch synchronous method." Local phase characteristics of the resynthesized speech are artificially regenerated by calculating the minimum-phase impulse response [13] from the square-root of each STRAIGHT-spectrum slice.

Current implementation of TANDEM-STRAIGHT extracts two types of excitation source information, fundamental frequency (F0) and aperiodicity information. The F0 information tightly correlates with a very important and dominant perceptual attribute "pitch." The aperiodicity information correlates with a delicate timbre-related attribute called "voice quality." Current implementation of this aperiodicity information is based on the power ratio of periodic component and random component in each frequency band (one octave wide). This information is summarized using a sigmoidal approximation with two parameters: boundary frequency and slope of transition [14].

2.4. Temporally variable multi-aspect morphing

Morphing based on STRAIGHT introduced a powerful means for manipulating singing voices, although it was originally designed to promote exploratory research of speech perception and production [15]. Reformulation of morphing algorithms based on interpolation/extrapolation of logarithmic transformation of derivatives of parameter mapping functions and exponential inversion enabled temporally variable multi-aspect morphing and made it more flexible and robust [2].

However, current excitation source representations are not rich enough to replicate realistic impressions of strong vocal expressions in singing voices. This makes this flexible morphing framework less effective for singing applications, and motivates the extension introduced in the following sections.

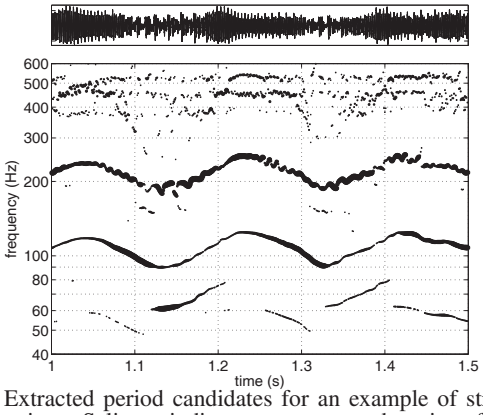

Fig. 1. Extracted period candidates for an example of strong vocal expressions. Salience indices are represented as size of the data dots. Plot shows an excerpted vowel part /a/ of a test performance by a male player of the Japanese traditional theatrical art "kyougen" reported in reference [20].

3. EXTENSION OF SOURCE REPRESENTATIONS

Rapid variation in timing and strength of excitation events is a typical characteristic of strong expressive voices. Such variation sometimes shows sub-harmonic (diplophonic) behavior. Three algorithms are prepared for analyzing such behavior: excitation structure extractor (XSX) [16], interval analysis of fundamental component [17] and a temporally stable instantaneous frequency representation [18]. The last two extensions are original in this article. Note that the following procedures assume high-quality recording (high SNR and wide frequency range), since post-processing of recorded singing voice is a primary target application.

3.1. XSX, multiple specialized periodicity detectors

An exhaustive periodicity detector, XSX, which detects all possible periodic candidates with periodicity salience indices, is designed by the following procedures. XSX was found to detect diplophonia and other complex excitation structures in Noh (Japanese traditional theatrical performance) singing voice [19] and pathological voices [16].

Dividing $P_T(\omega)$, a power spectrum of periodic signals, by $P_{ST}(\omega)$, its spectral envelope, leaves periodicity information and bias term. By removing bias and weighting to select base-band harmonic components, Fourier transform of the periodic information yields a salient peak at fundamental period, T_0. This yields a periodicity detector tuned to $T_0 = 1/f_0$.

Since no prior information about F0 is available, periodicity detectors spanning possible F0 range are allocated equidistantly on the logarithmic time (interval) axis, for example three detectors for each octave. Output of each detector is shaped so that the integrated detector output yields a constant value as a salience index [5].

Figure 1 shows an example of XSX analysis of strong vocal expressions by a male player of the Japanese traditional theatrical art "kyougen" reported in reference [20]. In this example, three types of periodic structures are visible: fundamental frequency (around 200 Hz), vibrato (about 5 Hz) and frequency and amplitude modulation (around 100 Hz, 70 Hz and 55 Hz) of the fundamental period and strength.

3.2. Interval analysis of fundamental component

Zero frequency filtering [21] is a simple and very fast algorithm for extracting prominent excitation events in voiced sounds. It essen-

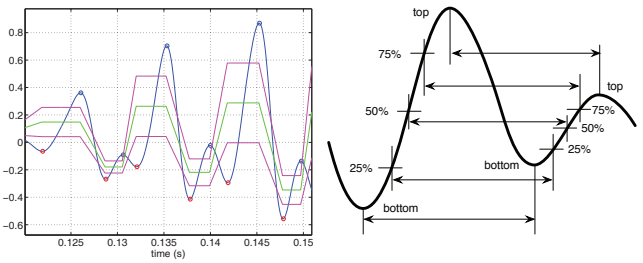

Fig. 2. (Left plot) Smoothed waveform and reference base lines for interval thresholding. (Green line) 50% level, (Mazenta lines) 25% and 75% levels. Peaks and dips are marked using small circles. (Right plot) Schematic illustration of selected intervals.

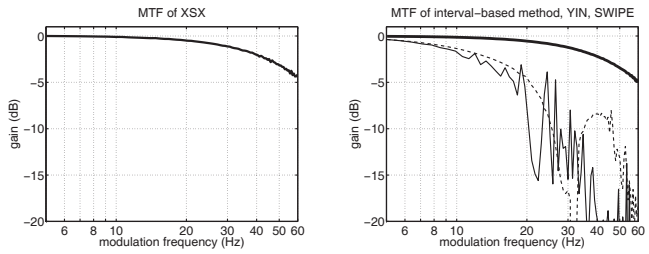

Fig. 3. Modulation transfer function to F0 frequency modulation. (Left plot) XSX. (Right plot) Thick line represents the proposed method, dashed line represents YIN and thin line represents SWIPE′

tially detects zero-crossing points of the fundamental component. Measuring intervals of multiple slice levels of the fundamental component provides means to evaluate the fundamental period and its salience of periodicity at the same time, with finer temporal resolution than other instantaneous frequency-based methods [17].

This method is extended to make it compatible with XSX by re-engineering its design. Fundamental component extraction and multi-level waveform thresholding require the windowing function to have high attenuation of sidelobe levels and asymptotic sidelobe decay rate steeper than -12 dB/oct. Usual windowing functions [22] do not meet these requirements. One practical selection is to use one of the reported Nuttall windows [23] with four cosine terms and sidelobe decay rate of -18 dB/oct. (Note that the selected window is not the commonly known "Nuttall window." Coefficients for zeroth through third cosine are 0.355768, 0.487396, 0.144232 and 0.012604, respectively. Refer to item 12 of Table II in [23].)

Figure 2 illustrates an example of multi-level thresholding (this time, five levels). By representing the threshold type by a variable K, where $K \in \{(\text{top}), (75\%), (50\%), (25\%), (\text{btm})\}$, the fundamental frequency is defined as an average of fundamental frequencies of all types of thresholding. Let M represent the set of thresholding types. Then, fundamental frequency $f_0(t)$, is calculated using the following equation:

$$f_0(t) = \frac{1}{n(M)} \sum_{K \in M} f^K(t), \tag{4}$$

where $f^K(t)$ represents the interpolated fundamental frequency by the type-K thresholding. Also, $n(M)$ represents the cardinal number of a set M and $f^K(t)$ is calculated using the following equation.

$$f^K(t) = f^K_{k_K} + \frac{t - l^K_k}{l^K_{k_K+1} - l^K_{k_K}} (f^K_{k_K+1} - f^K_{k_K}), \tag{5}$$

where $l^K_k = (t^K_k + t^K_{k+1})/2$ represents the nominal location of the calculated fundamental frequency $f^K_k = 1/(t^K_{k+1} - t^K_k)$ for type-K thresholding. t^K_k represents the type-K thresholding temporal location of the k-th positive-slope segment of the filtered waveform.

An F0 independent measure of periodicity salience is defined using standard deviation of logarithmic fundamental frequencies of different thresholding.

$$L = \exp\left(-\left(\frac{12^2}{n(M)} \sum_{K \in M} (\log_2(f^K(t)) - \log_2(f_0(t)))^2\right)^{\frac{1}{4}}\right), \tag{6}$$

where a constant "12" is used with musical applications in mind. Simulation tests using periodic signal plus white Gaussian noise with different SNR illustrated that the salience index is a good indicator of effective SNR, which cannot be directly observable.

Similar to XSX, no prior information about F0 is available for designing the necessary windowing function for suppressing harmonic components other than the fundamental one. In the current implementation, six windowing functions are prepared for each octave, assigning effective cut-off frequencies equidistantly on the log-frequency axis, and designed to cover 32 Hz to 1,000 Hz F0 range. Periodicity index L of each filter output is used to locate the best cut-off frequency. Using down-sampling and FFT-based fast convolution, this procedure runs three-times faster than real-time. (Everything is written in Matlab m-code, MacBook Pro, with 2.8 GHz Intel Core i7.) This speed makes it possible to use this procedure for screening out normal excitation segments before applying XSX, since XSX is computationally heavy (runs about ten-times slower than real-time on the same machine).

Figure 3 shows response of XSX and an interval-based method for synthetic signal with frequency-modulated F0. The center F0 is set to 200 Hz and modulation frequency spanned from 4 Hz to 64 Hz. The modulation depth was 100 cent (5.9463% of F0) peak-to-peak. Note that both methods can track 60 Hz modulation frequency and they outperform existing F0 extractors, such as YIN [24] and SWIPE′ [25]. (Default parameter setting of each method was used for executing YIN and SWIPE′.)

3.3. Temporally stable instantaneous frequency

Power weighted average of instantaneous frequencies is known to reduce annoying spiky behavior between harmonic frequencies. Substituting stable power spectral representation mentioned before [4, 1] in this averaging process using Flanagan's equation [26] enabled us to suppress this behavior completely [18]. The procedure was introduced for cases where two harmonic components are involved. The proposed procedure is generalized to more than two interfering harmonic components for improving temporal resolution and for reducing background noise effects.

For example, power weighted average of instantaneous frequencies using four temporal locations separated by $T_0/4$ yields temporally stable power spectrum for periodic signals, even when the windowing function has wider frequency domain representation, which covers up to three harmonic components. Substituting Flanagan's equation for instantaneous frequency, averaged instantaneous frequency is simplified to the following form.

$$\bar{\omega}_i(t) = \frac{\sum_{k=0}^{3} \left(\Re[X(t_k)] \Im\left[\frac{dX(t_k)}{dt}\right] - \Im[X(t_k)] \Re\left[\frac{dX(t_k)}{dt}\right] \right)}{\sum_{k=0}^{3} |X(t_k)|^2}, \tag{7}$$

where $t_k = t + T_0(k - 3/2)/4$ represents analysis locations and $X(t)$ represents filtered signal using quadrature signal for the filter impulse response. The windowing function introduced in the previous section is used to design the envelope of the quadrature signal. Values in discrete STFT bins are interpreted as filter outputs having nominal frequencies ω as filter center frequencies. Mapping from ω to ω_i in STFT is used to refine the initial estimate of fundamental frequencies obtained by the procedures mentioned in previous sections.

4. PRELIMINARY TEST AND DISCUSSION

These procedures were used to analyze several examples excerpted from a comprehensive singing voice recording [20] of mainly traditional Japanese singing stored on 18 audio CDs. Analysis and synthesis tests indicated that the extended source representations are capable of replicating a wide range of singing expressions.

Morphed sound quality deteriorates when these rapid parameter variations are directly mixed. Parametric representation such as audio texture [27] is relevant for introducing strong expression manipulation for the extended morphing framework. Manipulated singing examples can be found in our demonstration site [28].

5. CONCLUSION

Extension of excitation source representations are introduced to a speech analysis, modification and resynthesis framework, TANDEM-STRAIGHT. Fine temporal resolution of the proposed methods were found capable of extracting rapid F0 variations, which is typically found in expressive singing voice. Integration of these extensions to the exploratory environment based on the extended morphing algorithm is underway.

6. REFERENCES

[1] 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," *ICASSP2008*, pp. 3933–3936, 2008.

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

[3] 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.

[4] 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].

[5] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *SADHANA*, vol. 36, no. 5, 713–728, 2011.

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

[7] 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].

[8] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[9] B. H. Story, "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," *J. Acoust. Soc. Am.*, vol. 26, no. 1, pp. 327–335, 2008.

[10] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 505–519, 1995.

[11] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453 – 467, 1990.

[12] V. Villavicencio, A. R obel, and X. Rodet, "Applying improved spectral modeling for high quality voice conversion," *ICASSP2009*, pp. 4285–4288, 2009.

[13] A. V. Oppenheim and R. W. Shafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[14] H. Kawahara, M. Morise, T. Takahashi, H. Banno, R. Nisimura, and T. Irino, "Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems," *Interspeech2010*, pp. 38–41, 2010.

[15] H. Kawahara, "STRAIGHT, exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustic Science & Technology*, vol. 27, no. 5, pp. 349–353, 2006.

[16] Y. Wada, M. Morise, R. Nisimura, T. Irino, and H. Kawahara, "Optimization of a multiple local periodicity detector for vocal excitation structure analysis," in *Proc. APSIPA2010*, pp. 518–521, 2010.

[17] M. Morise, H. Kawahara, and T. Nishiura, "Rapid F0 estimation for high-SNR speech based on fundamental component extraction," *Trans. IEICE*, vol. J93-D, no. 2, pp. 109–117, 2010, [in Japanese].

[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, 2011.

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

[20] I. Nakayama, "Vocal expressions in Japanese by singing a common verse," *J. Acoust. Soc. Jpn.*, vol. 59, no. 11, pp. 688–693, 2003, [in Japanese].

[21] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[22] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[23] A. H. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. Audio Speech and Signal Processing*, vol. 29, no. 1, pp. 84–91, 1981.

[24] A. de Chevengné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[25] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[26] J. L. Flanagan and R. M. Golden, "Phase vocoder," *The Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[27] D.P.W. Ellis, Z. Xiaohong, and J.H. McDermott, "Classifying soundtracks with audio texture features," *ICASSP2011*, pp. 5880 –5883, 2011.

[28] http://www.wakayama-u.ac.jp/ kawahara/ICASSP2012/