Vol.2014-MUS-102 No.14
2014/2/23

Realtime conversion of growl-type voice qualities based on modulation and approximate time-varying filtering driven by a non-linear oscillator: Formulation

Kawahara Hideki[1,a)] Mizobuchi Shohei[1,b)] Morise Masanori[2,c)] Sakakibara Ken-Ichi[3,d)]
Nisimura Ryuichi[1,e)] Irino Toshio[1,f)]

Abstract: A formulation of voice conversion to add growl-like voice qualities to singing voices is proposed based on our findings of features in such singing performances. The proposed method does not consist of any analysis and synthesis stage(s). A preliminary implementation using Matlab demonstrated that its throughput is faster than realtime. The proposed formulation provides not only post processing capabilities of rendering styles of existing performances to recorded materials but also realtime capabilities of adding growl-like voice qualities in live performances.

1. Introduction

Strong emotional expression in singing sometimes results in aperiodic voices. Voice aperiodicity also is found in many traditional or ethnic singing [1], [2]. They make performance rich and moving. Definitely, voice aperiodicity is an important aspect of singing voice research.

However, such singing style is (or possibly is) not an easily acquired skill for majority of enthusiasts in singing. It also is a challenging target to develop effectors with flexible manipulation of singing voice aperiodicity [3]. One reason for such difficulty is its complexity. Actually, aperiodic voices involve complex interaction of voicing organs not only vocal fold but several different supra-laryngeal structures [1]. Valid simulation of aperiodic voice production process needs detailed understanding of underlying acoustical, mechanical, physiological and neural mechanisms and well exceeds scope of this article.

Instead of building a detailed model, we introduce a simple pipeline signal processing architecture based on our observations on growl-like singing [4] for designing realtime voice converter from normal voices to growl-like aperiodic voices. The following section briefly introduces those findings and outlines design goal.

2. Growl-like voices: analysis study

In out report, singing performances with and without growl-like expression were analyzed using an F0 (fundamental frequency) extractor with high-temporal resolution [5] and an F0 adaptive spectral envelope estimation method (STRAIGHT [6], [7], later). The following three dominant features were found to contribute perception of growl-like voice quality. [4]

F0 modulation: feature Q The F0 trajectory of growl-like singing consists of very fast (about 70 Hz) frequency modulation. The modulation power at the peak modulation frequency in growl-like singing is sometimes 20 dB higher than that of normal voice.

Spectral modulation: feature F Spectrographic representation of spectral envelope of growl-like singing has a specific texture with vertical lines. It is caused by temporal modulation of spectral envelope and seems to synchronize with the frequency modulation of F0.

Spectral enhancement: feature E Long-term spectrum of growl-like singing has higher power around 2000 Hz and lower power around 6000 Hz. Third octave level difference is used to design equalization filter to remove this feature.

The feature symbols (**Q, F** and **E**) are defined similar to our report [4]. Paired comparison test using selective suppression of these features yielded the following generalized linear model (GLM).

$$y = 1.287Q + 4.968F + 1.558E, \qquad (1)$$

where the estimate y represents the logit conversion of the preference probability of growl-like perception. Feature symbols Q, F and E represents difference of feature suppression between paired stimuli. (The value "1" represents existence of the feature and "0" represents absence. Therefore, the value of these independent variable is one element of the following set, $\{-1, 0, 1\}$.) All three features were found statistically significant.

[1] Wakayama University, Wakayama, Wakayama 640–8510, Japan
[2] University of Yamanashi, Kofu, Yamanashi 101–0062, Japan
[3] Health Science University of Hokkaido, Sapporo, Hokkaido 061–0293, Japan
[a)] kawahara@sys.wakayama-u.ac.jp
[b)] s155059@center.wakayama-u.ac.jp
[c)] mmorise@yamanashi.ac.jp
[d)] kis@hoku-iryo-u.ac.jp
[e)] nisimura@sys.wakayama-u.ac.jp
[f)] irino@sys.wakayama-u.ac.jp

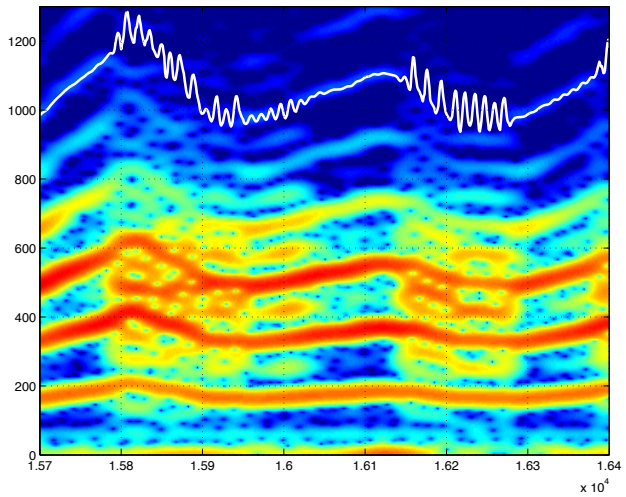

Fig. 1 Narrow band spectrogram of an excerpt of Noh voice with aperiodicity. White line represents sixth harmonic frequency ($6f_0(t)$).

2.1 Two databases

In addition to these analyses, two singing voice database were analyzed [8]. The first one is RWC music database for computer music research [9]. The database is a a collection of 100 audio CDs with full of copyright cleared contents for academic research purpose and individual sounds of musical instruments. Singing voices are stored as one category of musical instruments. It consists of three sopranos, three altos, three tenors, three baritones, three basses and three R&B singers' voices in various singing styles. The other database is a collection of very famous Japanese traditional vocal performance masters ' voices [10]. Some of the master performers are designated as the Japanese living national treasure. They are asked to sing a common verse in their traditional singing styles. They are also asked to sing Japanese five vowels. In addition to these compulsory recordings, they recorded traditional songs also. The voices are stored in eighteen audio CDs. It uses the B&K's 1/2" omni-directional condensor microphone (Type 4190), calibrated in pressure field. Both were sampled at 44100 Hz 16 bit format.

Similar coupled modulation of F0 trajectory and spectral envelope was also found in these databases. Close inspection of modulation behavior, these modulations were found to be phase-locked to the fundamental component, time to time.

Figure. 1 shows an example. It is an excerpt from a Noh performance with specific aperiodic behavior. In this magnified view, two groups of aperiodic period are shown. The white line in the figure shows the F0 trajectory. The fundamental frequency is multiplied by 6 and overlaid on the spectrogram. From 1.604 s to 1.62 s, fast F0 modulation frequency is $f_0/3$ and from 1.62 to 1.628 s, the modulation frequency is $f_0/2$ and they are phase-locked. The aperiodic period from 1.579 s to 1.595 s does not show phase-locking behavior and the modulation is chaotic. These behavior suggests that the modulation is caused by a coupled oscillation of supra-laryngeal structures [1].

This finding motivated the current implementation. "Phase-locked" nonlinear oscillator is the key component of the proposed procedure. Without phase-locking behavior, adding F0-frequency and spectral modulation to normal voices yielded a gargley voice.

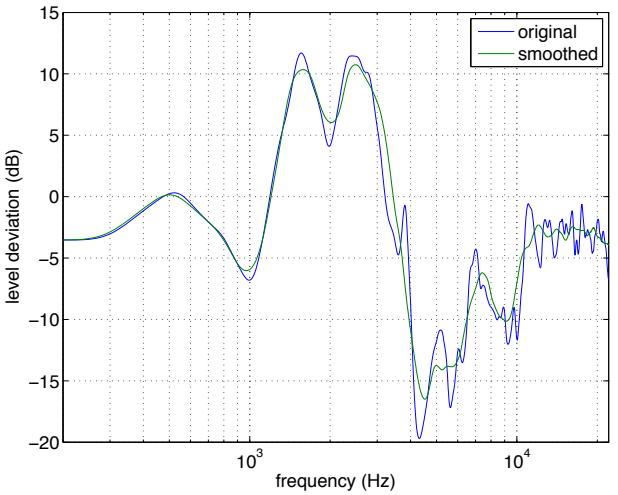

Fig. 2 Smoothed spectral envelope difference between growl-like voices and normal voices. The original long-term spectral difference is also shone.

3. Architecture and implementation

This section introduces architecture of the proposed method. It does not consist of analysis component. It only consists of oscillator, modulator and filters.

3.1 Temporal modulation: feature Q

Rapid F0 modulation found in growl-like singing can be approximately implemented by temporal axis modulation. Let $r_i(t)$ represent instantaneous ratio of fundamental frequency conversion from F0 $f_0^{IN}(t)$ of the input signal $x^{IN}(t)$ to F0 $f_0^{MOD}(t)$ of the output signal $x^{OUT}(t)$.

$$f_0^{MOD}(t) = r_i(t) f_0^{IN}(t) \tag{2}$$

Reading input signal using a modulated time axis converts apparent instantaneous frequency. This process is represented by the following equation.

$$x^{OUT}(t) = x^{IN}\left(\int_0^t \frac{1}{r_i(\tau)} d\tau\right) \tag{3}$$

For discrete time signals, piecewise linear interpolation provides the simplest implementation. This procedure is compatible with realtime processing by introducing a short buffer (shorter than 10 ms) and feedback control of its length.

This procedure does not model actual F0 modulation, because this procedure modulates spectral envelope of the input signal at the same time, while in actual speech production process, only source information is modulated. (Again, this explanation is also approximation when taking into account of nonlinear source filter interaction [11].) However, this approximation is close enough because the modulation depth of this fast variation does not exceed several semitones.

3.2 Approximate time-varying filter: feature F

The level equalization (feature **E**) can be implemented by using a FIR filter and will not be discussed in detail here. Figure 2 shows smoothed spectral difference with the original spectral difference.

Physically relevant implementation of the temporal variation of spectral envelope is to simulate wave propagation in the vocal tract with periodically changing area around supra-laryngeal structures. Wave propagation in a one-dimensional vocal tract model is accurately represented by the lattice filter architecture [12] and the reflection coefficients correspond to PARCOR coefficients [13] with relevant equalization procedures [14]. These physically well grounded parameters have relatively better interpolation behavior when parameters are temporally variable. However, this approach requires careful preparation of preprocessing procedures and relevant decision of model order. It also has difficulty when the sampling rate of the signal exceeds 8 kHz because validity of one dimensional vocal tract model is not assured in higher frequency range [15]. It also requires usually a long buffer length (about 30 ms or more) for calculating autocorrelation of the signal for estimating PARCOR coefficients. This buffer length introduces intrinsic delay and is not desirable for realtime processing. Appendix briefly discusses this alternative approach.

Instead of using physically relevant filter, a FFT-based approximate time-varying filter is introduced. Before introducing it, a brief review of FFT-based efficient implementation of (general) FIR filter is presented.

3.2.1 FFT-based convolution

FFT-based convolution is circular convolution. Let N represent the FFT buffer length and M and K represent the lengths of the signals to be convolved. FFT-based convolution provides same result when the condition $M + K < N$ is satisfied [16].

When the impulse response of a time-invariant filter $h[n]$ is a finite length sequence of length M, arbitrarily long sequence $x[n]$ can be filtered using FFT-based convolution by subdividing $x[n]$ into a set of subsequences $s^{(k)}[n]$ of length K using the following equation.

$$x[n] = \sum_{k=-\infty}^{\infty} s^{(k)}[n] = \sum_{k=-\infty}^{\infty} x_1^{(k)}[n] w[n - kL], \qquad (4)$$

where L represents frame shift and $w[n]$ represent the weighting sequence of length K. The k-th part $x_1^{(k)}[n]$ of the original sequence starts from $n = kL$ and its length is K. The weighting sequence $w[n]$ has to satisfy the following condition for all n.

$$\sum_{k=-\infty}^{\infty} w[n - kL] = 1 \qquad (5)$$

Arbitrary many weighting sequences satisfy this condition. The simplest example is $w_1[n] = 1, n = 0, 1, \ldots, K - 1$. This subdivides the original signal into non-overlapping consecutive sequences. The other common subdivision is to use $w_2[n]$ defined by the following equation.

$$w_2[n] = 0.5 - 0.5 \cos\left(\frac{2n\pi}{2L}\right), \qquad (6)$$

where the length is set $K = 2L+1$. This is 50% overlapping Hann window arrangement.

3.2.2 Time-varying filtering by FFT-based approximation

When the filter impulse response is time invariant, FFT-based convolution based on subdivision using $w_1[n]$ and subdivision using $w_2[n]$ are equivalent and the results are independent of the size of subdivision K and frame shift L.

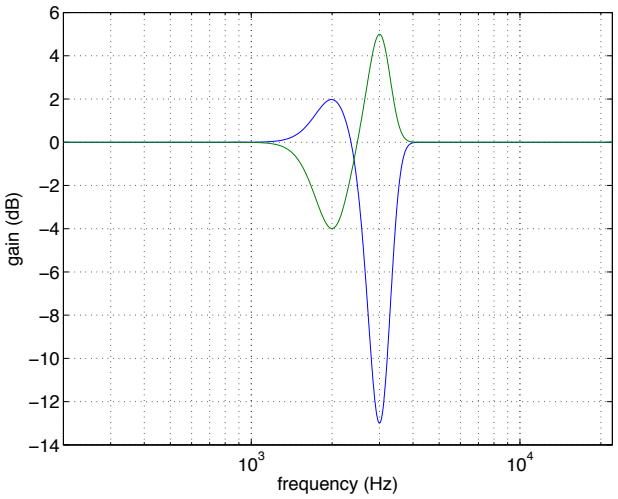

Fig. 3 Example of the spectral variation model. Blue line represents $R_1(f)$ and green line represents $R_12(f)$.

When the filter impulse response is changing temporally, FFT-based convolution based on subdivision using $w_1[n]$ introduces significant artifacts. The artifacts due to subdivision are small when using $w_2[n]$

The size of frame shift L depends on dynamic behavior of spectral envelop and the effective length M of the impulse response depends on the time varying spectral shape. They are application dependent design parameters.

In our current implementation, temporally varying spectral shape component is modeled using the following equation, two composite functions $R_1(f)$ and $R_2(f)$ consisting of Gaussian shapes. They represents spectral level variations around 3000 Hz and 2000 Hz changing in opposite direction.

$$R_1(f) = a_{p1} \exp\left(-\frac{(f - f_{p1})^2}{\sigma_{p1}^2}\right) - a_{d1} \exp\left(-\frac{(f - f_{d1})^2}{\sigma_{d1}^2}\right) \qquad (7)$$

$$R_2(f) = a_{p2} \exp\left(-\frac{(f - f_{p2})^2}{\sigma_{p2}^2}\right) - a_{d2} \exp\left(-\frac{(f - f_{d2})^2}{\sigma_{d2}^2}\right), \qquad (8)$$

where f_{p1} and f_{p2} represent two peak frequencies, σ_{p1}^2 and σ_{p2}^2 represent corresponding peak widths and a_{p1} and a_{p2} represent their peak levels respectively. Similarly, f_{d1} and f_{d2} represent two dip frequencies, σ_{d1}^2 and σ_{d2}^2 represent corresponding dip widths and a_{d1} and a_{d2} represent their dip levels respectively. Temporally variable filter shape $R_m(f, t)$ is defined using a mixing factor $r_m(t)$.

$$R_m(f, t) = r_m(t) R_1(f) + (1 - r_m(t)) R_2(f) \qquad (9)$$

Figure 3 shows example shapes of $R_1(f)$ and $R_2(f)$. Figure 4 shows corresponding minimum phase impulse responses [16] using logarithmic vertical axis. (Parameters are as follows: $f_{p1} = f_{d2} = 2000$ Hz, $f_{p2} = f_{d1} = 3000$ Hz, $\sigma_{p1} = \sigma_{p2} = \sigma_{d1} = \sigma_{d2} = 400$ Hz, $a_{p1} = 2$ dB, $a_{d1} = 13$ dB, $a_{p2} = 5$ dB, $a_{d2} = 4$ dB) It illustrates that the responses effectively vanish within 3 ms and it provides the upper bound of the response length M. (For 44100 Hz

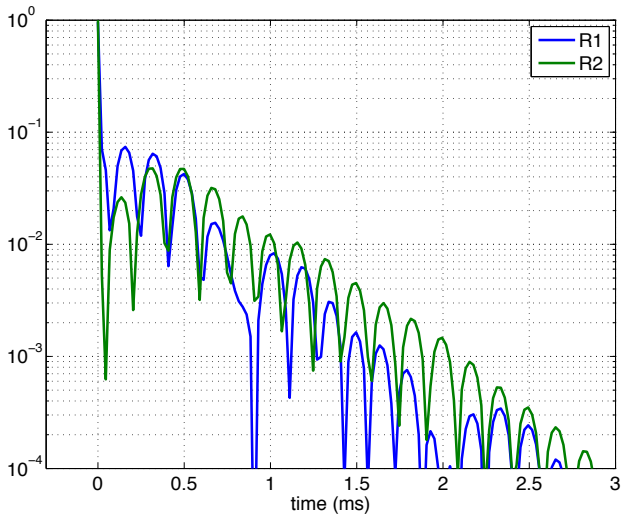

Fig. 4 Log-magnitude plot of impulse responses of $R_1(f)$ and $R_12(f)$.

sampling, $M \leq 133$). Since the dominant F0 frequency modulation frequency is about 70 Hz, relevant frame update rate is 2 ms ($L \approx 88$). The length of the subdivided segment is 177 when using $w_2[n]$. The minimum best FFT buffer length N_{min} is calculated by the following equation, since performance of FFT is best when the buffer length is 2^Z (Z: natural number).

$$N_{min} = 2^{\lfloor \log_2(M+K+1) \rfloor}. \tag{10}$$

For the current example, $N_{min} = 512$.

3.3 Test using sinusoidal F0 modulation

All necessary procedures to implement features (**Q, F** and **E**) found in growl-like performance are implemented. A preliminary test was conducted using a 70 Hz sinusoidal signal is used to simulate the modulation. Combining all features successfully added growl-like impression to normal singing and will be reported elsewhere [17]. However, careful listening revealed that the converted voices sounded rather wet voice (gargley voice) than growl-like. This finding motivated following investigations on phase-locked oscillation of nonlinear oscillators.

4. Nonlinear oscillator and coupling

The following equation provides sinusoidal oscillation of variable x without severe distortion found in the Van der Pol oscillator.

$$\frac{d^2x}{dt^2} + \varepsilon\left(\left(\frac{dx}{dt}\right)^2 + x^2 - 1\right)\frac{dx}{dt} + x = 0, \tag{11}$$

where ε represents contribution of speed dependent factor similar to the Van der Pol oscillator. It is convenient to define additional variable $y = \frac{dx}{dt}$ and yields the following equation. A set of variables x and y represents the state of the oscillator.

$$\frac{dx}{dt} = y \tag{12}$$

$$\frac{dy}{dt} = \varepsilon\left(1 - x^2 - y^2\right)y - x \tag{13}$$

Using this equation and by introducing a signal $s(t)$ to control the status of oscillation yields the following equation.

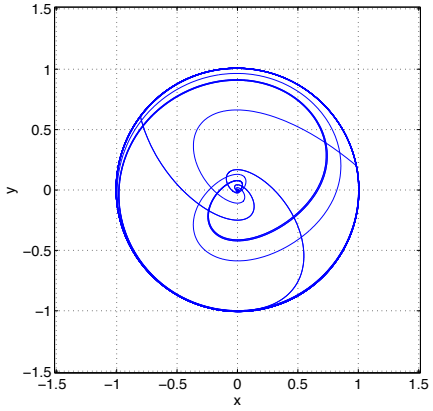

Fig. 5 Trajectory of controlled oscillation in the state space.

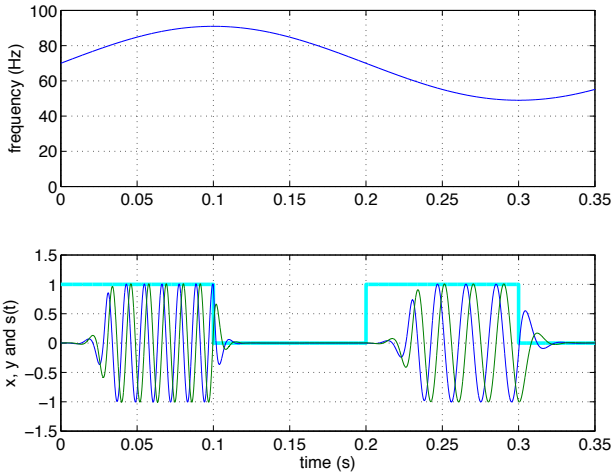

Fig. 6 Controlled oscillation and its control functions. Upper plot shows the oscillation frequency $f_m(t)$. Thick cyan line in the lower plot represents ON/OFF control signal $s(t)$.

$$\frac{dx}{dt} = y \tag{14}$$

$$\frac{dy}{dt} = s(t)\varepsilon\left(1 - x^2 - y^2\right)y - x - (1 - s(t))\varepsilon y, \tag{15}$$

when $s(t) = 1$, the second equation is reduced to Eq. (13). When $s(t) = 0$, the second equation is reduced to a dumping oscillator. Then, by using the chain rule, the frequency of oscillation $f_m(t)$ (here, temporally variable oscillation frequency is assumed) is directly controlled by using $\omega_m(t) = 2\pi f_m(t)$ in the following equation.

$$\frac{dx}{dt} = \omega_m(t)y \tag{16}$$

$$\frac{dy}{dt} = \omega_m(t)\left(s(t)\varepsilon\left(1 - x^2 - y^2\right)y - x - (1 - s(t))\varepsilon y\right) + F(t), \tag{17}$$

where $F(t)$ represents the external force. In the proposed procedure, this external force corresponds to the fundamental component of the input signal.

Figure 5 shows the state trajectory of the oscillator. Figure 5 illustrates that the attractor of the oscillation is circular and leads to sinusoidal oscillation. Figure 6 illustrates frequency and ON/OFF control by the control functions. The coefficient $\varepsilon = 1$ is used in this example. Note that response to the ON/OFF control is

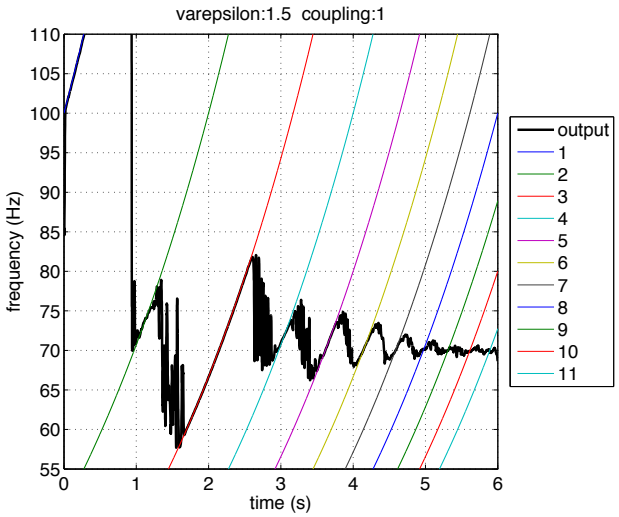

Fig. 7 Phase locking to input signal. Horizontal axis represents instantaneous fundamental frequency of input square wave. Vertical axis represents instantaneous frequency of the oscillator.

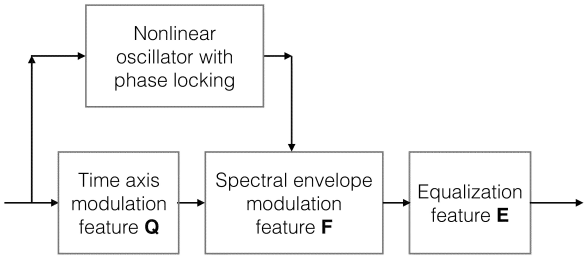

Fig. 8 Schematic diagram of the proposed method.

smooth.

Figure 7 shows an example of phase locking behavior of this oscillator. Input is a chirp square wave. The instantaneous fundamental frequency of the signal starts from 100 Hz and log-linearly rises up to 800 Hz in six seconds. The colored lines represent the chirp frequency divided by the integers shown in the legend of the figure. It indicates that this oscillator is tend to be phase locked to the 1/3 of the input signal's F0.

4.1 Application to growl-like converter

This nonlinear oscillator was used to replace the sinusoidal oscillator for driving time-axis modulator (feature **Q**) and approximate time-varying filter (feature **F**). Figure 8 shows schematic diagram of the proposed method.

A prototype system was implemented using Matlab. By replacing independent sinusoidal oscillator with a nonlinear oscillator, problematic gargley voice timbre was somewhat reduced but still remains. It may suggest need of model refinement of temporal spectral variations.

5. Conclusion

A simple and potentially realtime voice conversion method to add growl-like voicing style on normal singing is formulated. It consists of temporal modulation, approximate time-varying filter and nonlinear oscillator coupled to the fundamental component of the input voice. Because of this simple and straightforward architecture, the proposed prototype runs faster than realtime even using Matlab. However, lack of audio output function that is compatible with realtime processing, it is difficult to implement the proposed algorithm using Matlab only. Implementation using other language is currently undertaken.

Acknowledgments This work was partly supported by Grants-in-Aid for Scientific Research category (B)24300073 and Exploratory Research 24650085. It is also by Wakayama University.

References

[1] Sakakibara, K., Fuks, H., Imagawa, N. and Tayama, N.: Growl voice in ethnic and Pop styles, *Proc. Int. Symp. on Musical Acoustics* (2004).
[2] Fujimura, O., Honda, K., Kawahara, H., Konparu, Y., Morise, M. and Williams, J.: Noh Voice Quality, *J. Logopedics Phoniatrics Vocology*, Vol. 34, No. 4, pp. 157–170 (2009).
[3] Bonada, J. and Blaauw, M.: Generation of growl-type voice qualities by spectral morphing, *Proc. ICASSP2013*, pp. 6910–6914 (2013).
[4] Kawahara, H., Morise, M. and Sakakibara, K.: Interference-free observation of temporal and spectral features in "shout" singing voices and their perceptual roles, *Proc. SMAC2013* (Bresin, R. and Askenfelt, A., eds.), KTH Royal Institute of Technology, pp. 256–263 (2013).
[5] Kawahara, H., Morise, M., Nisimura, R. and Irino, T.: Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution, *ICASSP2013*, pp. 6797–6801 (2013).
[6] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, *ICASSP2008*, pp. 3933–3936 (2008).
[7] Kawahara, H. and Morise, M.: Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework, *SADHANA*, Vol. 36, No. 5, pp. 713–722 (2011).
[8] Kawahara, H., Morise, M. and Sakakibara, K.: Temporally fine F0 extractor applied for frequency modulation power spectral analysis of singing voice, *Proc. MAVEBA 2013* (Manfredi, C., ed.), Firenze University Press, pp. 125–128 (2013).
[9] Goto, M.: Development of the RWC Music Database, *Proc. ICA 2004*, pp. I–553–556 (2004).
[10] Nakayama, I.: Comparative studies on vocal expression in Japanese traditional and western classical-style singing, using a common verse, *Proc. ICA 2004*, pp. 1295–1296 (2004).
[11] Titze, I. R.: Nonlinear source–filter coupling in phonation: Theory, *J. Acoust. Soc. Am.*, Vol. 123, No. 5, pp. 2733–2749 (online), DOI: 10.1121/1.2832337 (2008).
[12] Wakita, H.: Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art, *Acoustics, Speech and Signal Processing, IEEE Transactions on*, Vol. 27, No. 3, pp. 281 – 285 (1979).
[13] Itakura, F. and Saito, S.: Speech analysis-synthesis system based on the partial autocorrelation coefficients, *Proc. Conv. Acoust. Soc. Japan*, No. 2-2-6, pp. 199–200 (1969).
[14] Arakawa, A., Uchimura, Y., Banno, H., Itakura, F. and Kawahara, H.: High quality voice manipulation method based on the vocal tract area function obtained from sub-band LSP of straight spectrum, *Proc. ICASSP 2010*, pp. 4834–4837 (2010).
[15] Ternström, S. O.: Hi-Fi voice: observations on the distribution of energy in the singing voice spectrum above 5 kHz, *Proc. Acoustics'08 Paris*, pp. 3171–3176 (2008).
[16] Oppenheim, A. V. and Shafer, R. W.: *Discrete-Time Signal Processing*, Prentice-Hall,, Englewood Cliffs, NJ (1987).
[17] Mizobuchi, S., Nisimura, R., Irino, T. and Kawahara, H.: Analysis and reproduction of growl type singing focused on temporal variation of spectral shape, *Proc. Spring Meeting Acoust. Soc. Japan*, No. 2-Q5-20 (2014).
[18] Kawahara, H.: Matlab realtime speech tools, Wakayama University (online), available from ⟨http://www.wakayama-u.ac.jp/%7Ekawahara/MatlabRealtimeSpeechTools/⟩ (accessed 29/January/2014).

Appendix

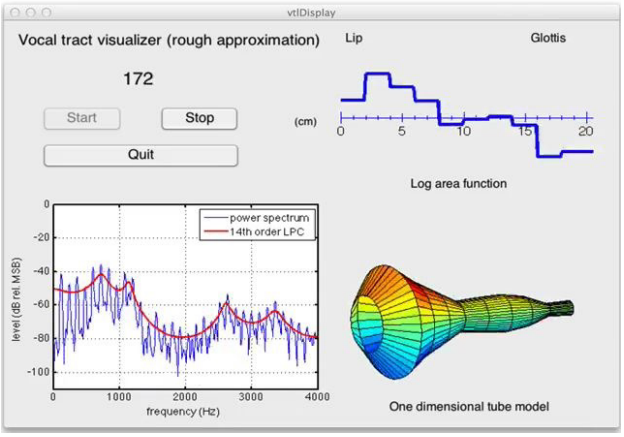

Fig. A·1 GUI of realtime vocal tract visualizer implemented by Matlab

A.1 Realtime vocal tract visualizer

Recent audio input object of Matlab `audiorecorder` introduced data acquisition capability from the object while it is running. This functionality enabled implementation of realtime tools solely using Matlab. Figure A·1 shows a snapshot of a GUI which can visualize speaker's vocal tract area function and related information in realtime. By modulating, for example, a part of the displayed log-area function corresponding to laryngeal area, more authentic growl-like conversion can be implemented. This tool and other useful tools are linked to the first author's web page. The direct link is [18].