Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and F0 extractor evaluation

Hideki Kawahara*, Ken-Ichi Sakakibara†, Hideki Banno‡, Masanori Morise§ Tomoki Toda¶ and Toshio Irino*
* Wakayama University, Wakayama, Japan
E-mail: {kawahara,Irino}@sys.wakayama-u.ac.jp
† Health Sciences University of Hokkaido, Sapporo, Japan
E-mail: kis@hoku-iryo-u.ac.jp
‡ Graduate School of Science and Technology, Meijo University, Nagoya, Japan
E-mail: banno@meijo-u.ac.jp
§ Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu, Japan
E-mail: mmorise@yamanashi.ac.jp
¶ Information Technology Center, Nagoya University, Nagoya, Japan
E-mail: tomoki@icts.nagoya-u.ac.jp

***Abstract*—A closed-form representation of anti-aliased L-F model is derived for a LPF function family based on cosine series. The Matlab based implementation of the derived form provides virtually aliasing-free source signal, which is applicable to speech synthesis and F0 extractor evaluation. This aliasing-free representation is also suitable for testing perceptual effects of wave shape parameters in the L-F model, since possible artifacts caused by spurious component are completely removed. A post processing procedure for fine tuning spectral shape is also introduced. An interactive tool for investigating speech production model parameters is designed using this Matlab implementation and will be made freely available.**

I. Introduction

The L-F model [1], [2] of voice excitation source, which models air flow at glottis and radiation effect, has been applied to various fields in speech science and applications. Even though the actual vocal fold vibration [3] has far more complex details, the L-F model provides a relevant abstraction of essential aspects. It is advantageous to use it for the excitation source signal for speech synthesis [4], [5], [6], because it allows us to make use of large amount of findings in physiological [2], [7] as well as perceptual [8], [9] correlates of model parameters.

However, for applying the L-F model to speech synthesis, there is a fundamental issue to be considered. The L-F model is defined as a piecewise exponential function and consists of discontinuities in its derivative. These discontinuities introduce significant amount of spurious components caused by aliasing when sampled for discretization. This aliasing problem was already solved for an excitation source model based on a piecewise polynomial [10] by explicitly introducing anti-aliasing function for deriving closed-form representations. Unfortunately, the L-F model is represented as a piecewise exponential function and the procedure used in the reference [10] is not directly applicable.

This article introduces a closed-form representation of the L-F model applicable to speech synthesis and objective evaluation of F0 extractors. It also introduces an alternative implementation, which is easier and more flexible for practical applications and illustrates that this alternative representation is effectively identical to the closed-form solution.

II. Background

Voiced sounds are excited by the intermittent air flow caused by valving motion of vocal folds. When the glottis is open, the exhalation air flow goes through the opening. When the glottis is closed, the air flow is stopped. This valving action introduces discontinuities in higher order derivatives of the volume velocity of the air flow [11]. The L-F model [1] is defined to represent this phenomena and to take into account of the radiation characteristics from opening (for example, mouth opening).

The L-F model is uniquely determined by setting a specific value to each parameter, t_p, t_e, t_a, t_c and T_0 shown in Fig. 1. It is convenient to represent these parameters normalized by T_0. The values in Fig. 1 uses this normalization. These parameters have the following meanings.

- t_p location of the maximum volume velocity
- t_e location of the first contact of vocal folds
- t_a projection of the derivative of latter part at t_e on the time axis.
- t_c location of the complete closure

Figure 1 shows a slightly different definition of parameters of the L-F model. This modification is for making the definition compatible with the reference [7]. In this definition, timing of the complete closure t_c is not necessary to match the fundamental period T_0.

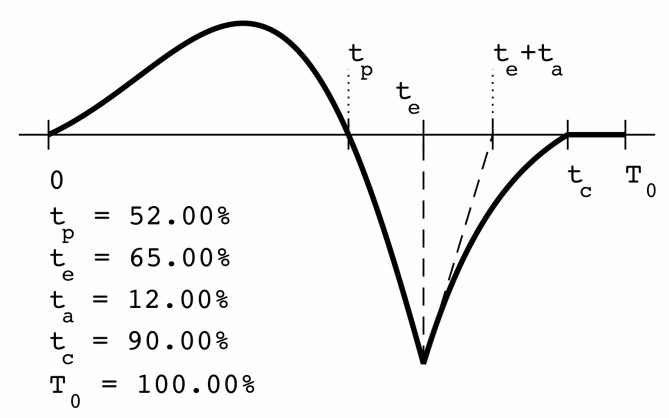

Fig. 1. Definition of L-F model parameters and waveform. This definition is extended to allow independent setting of T_0 and t_c to make it compatible to [7]

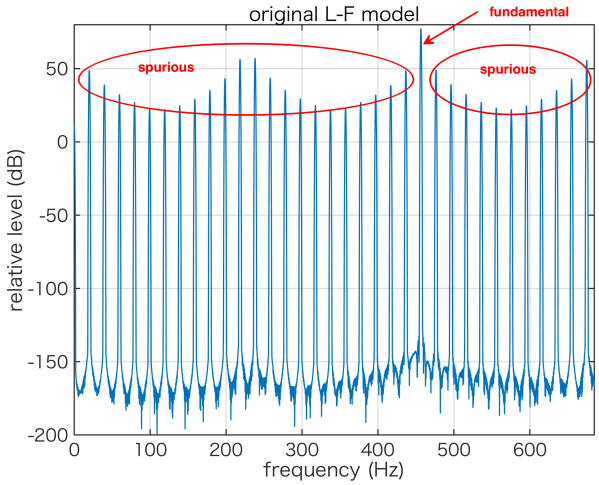

Fig. 2. Power spectrum example of the discretized L-F model using 8 kHz sampling frequency. The fundamental frequency is designed to make aliasing easily visible. The parameters of the L-F model of this example corresponds to "modal" voice quality [7].

Using these parameters, the L-F model is defined by the following equation.

$$E(t) = E_0 e^{\alpha t} \sin \omega_g t \qquad (t < t_e) \quad (1)$$

$$E(t) = \frac{-E_0}{\beta t_a}\left[e^{-\beta(t-t_e)} - e^{-\beta(t_c-t_e)}\right] \qquad (t_e \le t < t_c), \quad (2)$$

where E_0 represents a coefficient to adjust magnitude of the signal. The values of the growing/decaying parameters α and β, as well as the vibration speed parameter ω_g are calculated from the given set of temporal parameters.

Distribution of these parameters in different types of voice quality has been investigated by many researchers [2], [7] and methods for estimating these parameters have been studied [12], [13], [5], [14]. Also, perceptual effects of these parameters [15] have been investigated for decades.

A. Aliasing

The source signal generated by the L-F model consists of three discontinuities in its first order derivative. They are located at $t = 0, t = t_e$ and $t = t_c$. These discontinuities are not band-limited and consequently introduce spurious components caused by aliasing when sampled for discretization.

This aliasing effects are severe when the sampling frequency is relatively low. Figure 2 shows an example. It shows the magnified view of the power spectrum of a L-F model signal, which is sampled at 8000 Hz. It shows the power spectrum of the signal from 0 to $1.5 f_0$, where $f_0 = 8000/(17+12/23) \approx 456.58$ (Hz). The components other than F0 in the figure are caused by aliasing. The time window function used in this analysis is designed using the discrete prolate spheroidal wave function [16] to assure low side lobe levels (lower than -200 dB) and the best time-bandwidth product for finite support functions [17]. These spurious components are audible and degrade the synthesized speech sounds when they are used for speech synthesis.

B. Closed-form of band-limited source model

A piecewise polynomial model of the glottal source signal was proposed and derived three types of closed-form representations of its band-limited versions [10]. The formulation enabled continuous control of F0 of the discretized version of the source signal without introducing severe aliasing effects. This representation was used to test several F0 extractors and perceptual effects of aliasing were also evaluated [10]. Unfortunately, the glottal source model used in this literature is not directly comparable to the widely used L-F model. Deriving an explicit form of the band-limited version of the L-F model is the main contribution of this article. BLIT [18] also provide a systematic method to produce anti-aliased discrete signal from continuous time signals with discontinuities and can be applied to the L-F model. Our closed form representation allows more direct and flexible manipulation of parameters than BLIT-based implementation.

III. Closed-form representation of anti-aliased L-F model

A set of closed-form representations for constituent piecewise functions is derived for the L-F model and cosine series LPF functions [19]. The resulted representation only consists of complex exponentials and easily and efficiently implemented. Please refer to Appendix-A for details.

A. Anti-aliasing parameters

Let $E_b(t;\Theta)$ represent the band-limited version of the L-F model $E(t)$. The symbol Θ represents the set of parameters to specify the cosine series LPF function. It is defined as follows.

$$\Theta = \{T_w, w_0, \ldots, w_N\}, \qquad (3)$$

where T_w represents the half of the support length of the function $h(t)$ and $w_k, k = 0, \ldots, N$ represent coefficients of

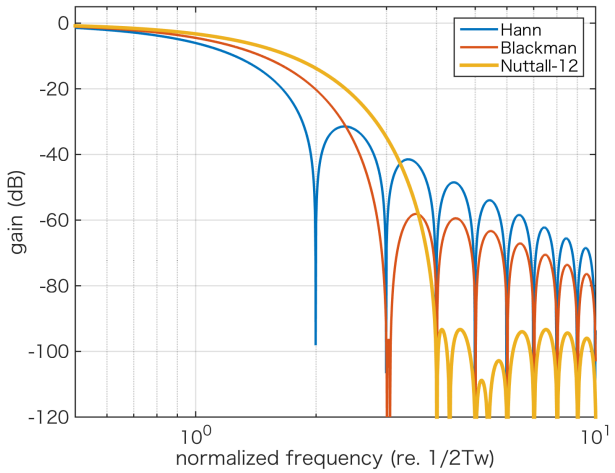

Fig. 3. Frequency responses of Hann, Blackman and Nuttall windowing functions. The frequency axis is normalized in terms of the windowing length $2T_w$.

the cosine series defined below.

$$h(t) = \sum_{k=0}^{N} w_k \cos\left(\frac{\pi k t}{T_w}\right), \quad \text{for} - T_w < t < T_w. \qquad (4)$$

Several useful set of parameters are given in [20], [19]. Hann, Blackman and Nuttall windows are in the set. Figure 3 shows frequency responses of them. The frequency axis is represented using the normalized frequency in terms of the total window length $2T_w$. For anti-aliasing purpose, more than 90 dB suppression of side lobes and their rapid decay (-18 dB/oct) of the Nuttall window is relevant. The coefficients of this Nuttall window $\{w_k\}_{k=0}^{3}$ are 0.355768, 0.487396, 0.144232, and 0.012604. This set is the 12-th item of the Table 1 of the reference [19]. The first zero of the spectral representation of the Nutall window is located at the normalized frequency $2/T_w$. This zero should be adjusted to $f_s/2$ when applying this window for anti-aliasing. Note that the parameter T_w is represented in terms of the normalized time axis, which is defined in terms of the fundamental period $T_0 = 1/f_0$. Consequently, the value of T_w for the Nuttall window on the normalized time axis is given by the following.

$$T_w = \frac{4 f_0}{f_s}. \qquad (5)$$

B. Anti-aliased waveform and spectral shape

Figure 4 and 5 show the original signal and its anti-aliased versions for modal voice and breathy voice. In these plots, the fundamental frequency is set $f_0 = 8000/(17 + 12/23) \approx 456.58$ (Hz). The anti-aliasing parameter T_w was determined according to the assumed sampling frequency and f_0 using (5). The figures show results for 8000, 16000, 44100, and 48000 Hz sampling frequencies.

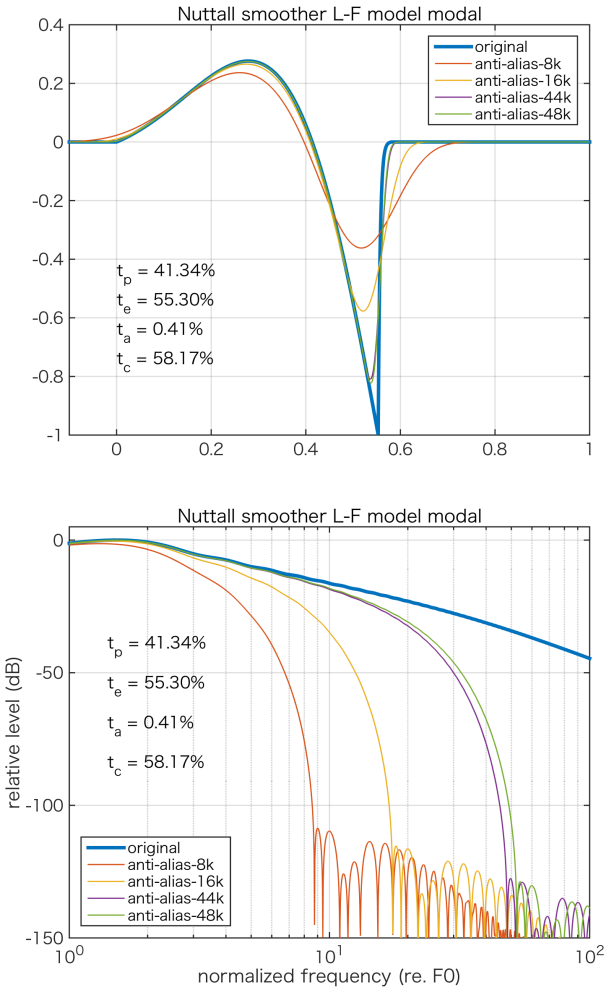

Fig. 4. Comparison of the original L-F model source signal and anti-aliased version. Upper plot shows waveforms and the bottom plot shows spectra.

Waveform plots in Figs. 4 and 5 show that lower sampling frequency makes waveform smoother. Spectrum plots in Figs. 4 and 5 show that side lobe levels (in this case the aliasing levels) are suppressed more than 100 dB. This shows that by using this anti-aliasing operation, spurious components shown in Fig. 2 effectively vanish.

C. Source signal with arbitrary F0 trajectories

This section introduces a procedure to generate the anti-aliased L-F signal for a time varying F0 trajectory, $f_0(t)$. Assume that the F0 trajectory $f_0(t)$ is a continuous function of class C^0. Then, this trajectory is subdivided into segments separated at points where the principal value of the following phase $\phi(t)$ has discontinuities.

$$\phi(t) = \arg\left(\exp\left(2\pi j \int_0^t f_0(\tau) d\tau\right)\right), \qquad (6)$$

where $\arg(x)$ is the function to yield the principal value of the phase of the complex number x.

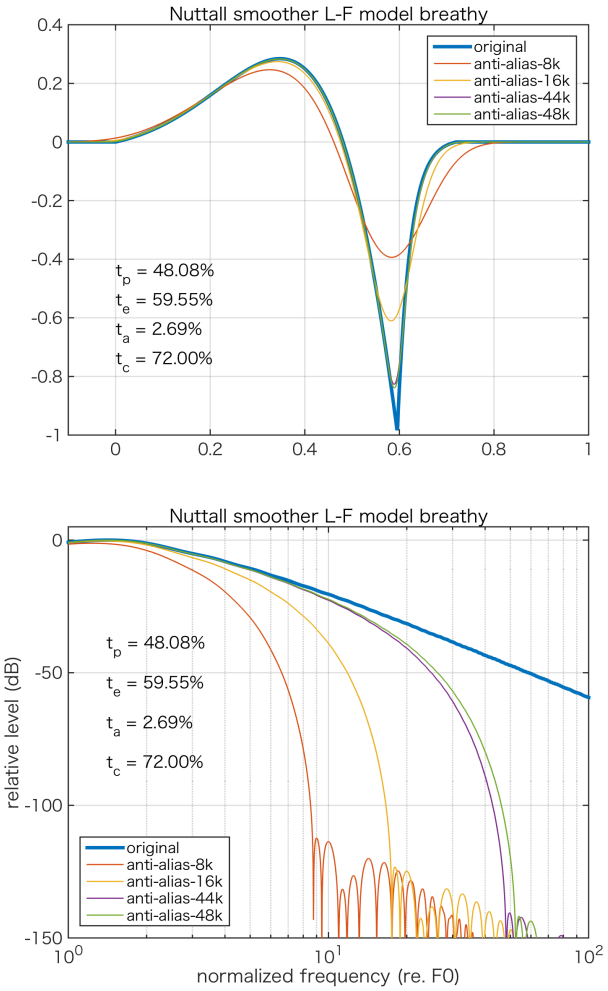

Fig. 5. Comparison of the original L-F model source signal and anti-aliased version. Upper plot shows waveforms and the bottom plot shows spectra.

The normalized time axis $\lambda_k(t)$ for the L-F model is given for each segment, for example k-th segment is represented using the subdivided phase function $\phi_k(t)$.

$$\lambda_k(t) = \frac{\phi_k(t) + \pi}{2\pi}. \tag{7}$$

Using this time axis, the anti-aliased source signal $E_{bv}(t, f_0(t); \Theta)$ for the time varying F0 trajectory $f_0(t)$ is given by the following equation.

$$E_{bv}(t, f_0(t); \Theta) = \sum_k f_0(t) E_b(\lambda_k(t); \Theta), \tag{8}$$

where $f_0(t)$ as the coefficient is used to fulfill the following condition.

$$\int_{-\infty}^{\infty} E_b(\lambda_k; \Theta) d\lambda_k = 0. \tag{9}$$

D. Matlab implementation

The proposed form is implemented using Matlab, an environment for scientific computation. In this implementation, the

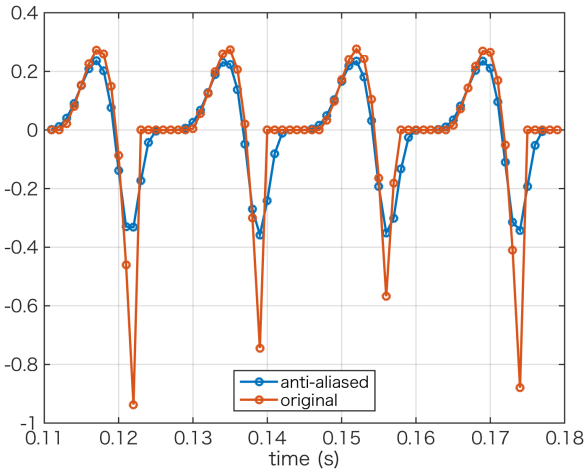

Fig. 6. Discretized source waveform of the original L-F model and the anti-aliased L-F model. Small circles on each line represent the discretized points. The parameter set for modal voice quality is used.

cumulative summation function `cumsum` is used for integration. It also should be noted that each subdivision has to be extended to both directions to take into account of the signal smearing caused by smoothing. The simplest way to extend is by redefining the subdivided normalized time axis $\lambda_{Ek}(t)$ as follows.

$$\lambda_{Ek}(t) = \lambda_{k-1}(t) - u(t_k - t) + \lambda_k(t) + \lambda_{k+1}(t) + u(t - t_{k+1}), \tag{10}$$

where t_k represents the starting point of the k-th segment and $u(t)$ represents the unit step function. It is practical to use this extended normalized time in $-0.5 < \lambda_{Ek}(t) < 1.5$.

E. Numerical example

A set of test signals are generated to illustrate effects of the closed form representation of the L-F model with closed-form anti-aliasing. The F0 trajectory for the test signal is defined below.

$$f_0(t) = \frac{8000}{17 + \frac{12}{23}} \cdot 2^{\left(\frac{1}{24}\sin(2\pi 5.5 t)\right)}, \tag{11}$$

where 5.5 defines the rate of vibrato, which depth is half semitone peak-to-peak. The center fundamental frequency is set similar to Fig. 2.

Figure 6 shows the discretized source signals, the directly discretized original L-F model and the discretized anti-aliased L-F model. The minimum peaks of the direct L-F model show variation in each cycle, while those of the anti-aliased signal do not show such behavior. The reproduced sounds of the original L-F model consists of clearly audible noise, which is caused by these variations. The reproduced sound of the anti-aliased signal sounds clean and consists of no noise.

Figure 7 shows the spectrograms of the generated signals. Nuttall window with 50 ms in length is used to analyze the signal. The display dynamic range is set 90 dB, since the

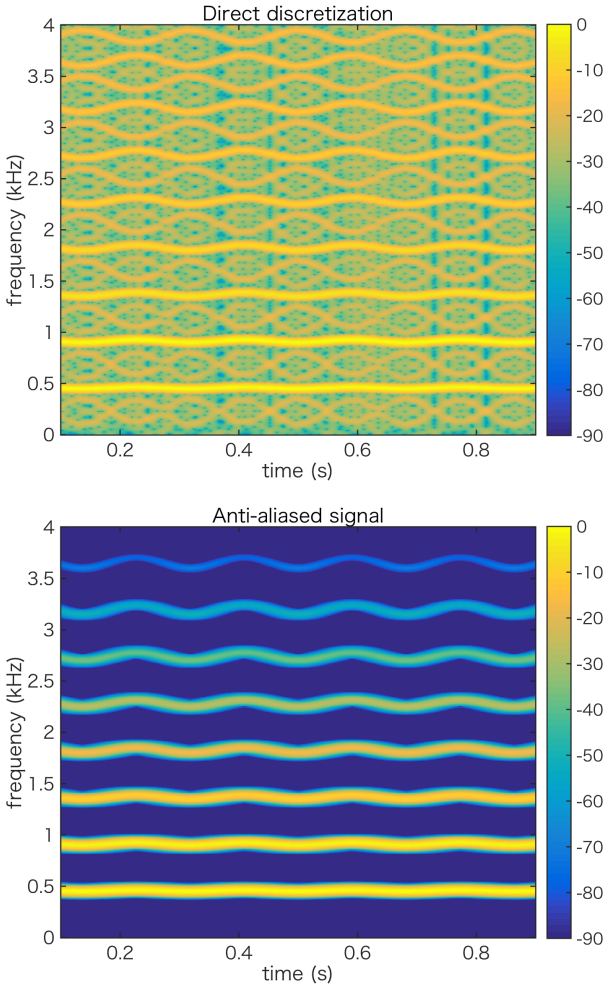

Fig. 7. Spectrograms of the discretized source signals. Note that strong spurious components are visible in the original L-F model.

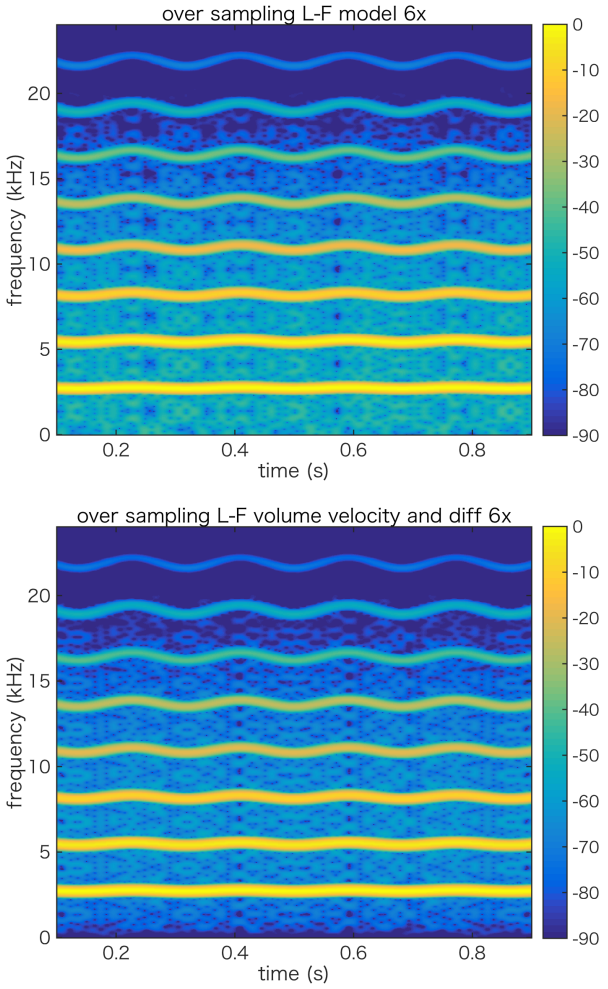

Fig. 8. Spectrograms using over-sampling practice. This time, 6 times over-sampling is used. The upper spectrogram shows direct application to the over-sampled L-F model. The lower spectrogram shows application to the over-sampled volume velocity based on the L-F model followed by filtering and down-sampling.

maximum side lobe level of this Nuttall window is -93.6 dB from the peak of the main lobe. The spectrogram of the anti-aliased signal does not show any trace of component other than harmonics. It illustrates that the proposed procedure effectively provides aliasing-free discretized L-F model signal. The spectrogram of the discretized original L-F model consists of full of spurious components other than harmonics. This illustrates that directly discretizing the L-F model waveform is potentially harmful for speech synthesis applications.

IV. Alternative Implementation

Over-sampling and anti-aliasing filtering on the over-sampled time domain is a common practice. Figure 8 shows spectrograms using this practice. The over-sampled signal is using 48000 Hz sampling frequency, six times over-sampling. The upper spectrogram is calculated from the over-sampled L-F model output directly. The lower spectrogram is calculated from the over-sampled volume velocity signal based on L-F model. Appendix-B provides a set of equations to represent the L-F model-based volume velocity. The anti-aliasing LPF is designed on 48000 Hz sampling system and applied to the over-sampled signal. Then, the anti-aliased L-F model signal is down-sampled. For the volume velocity signal, differentiation is applied to the down-sampled signal.

Comparison of Fig. 7 and Fig. 8, illustrates that levels of spurious components are substantially reduced by over-sampling, especially for the volume velocity signal. Figure 9 provides support of this observation quantitatively. It shows the temporally averaged spectrograms. It also consists of the result of the direct L-F model discretization.

The direct discretization introduces spurious level around -23 dB to the fundamental component. This spurious level is substantially reduced to about -50 dB using over-sampling and anti-aliasing filtering in the over-sampled domain. By using the volume velocity signal for over-sampling, the level is further reduced to about -70 dB. For the anti-aliased closed-form, the spurious level is around -120 dB. This is a substantial

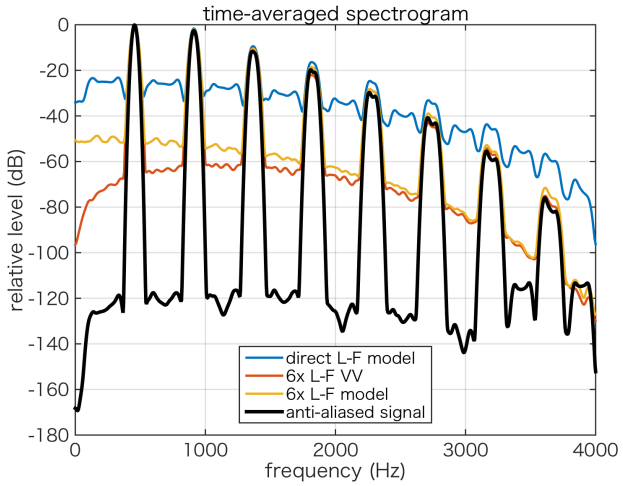

Fig. 9. Effects of over-sampling. Temporal average of spectrograms are shown. Note that the same Nuttall-based LPF designed in 8000 Hz sampling system is applied to the direct discretization of the L-F model.

reduction.

In this figure, the spectrogram for the anti-aliased signal is calculated using the self convolution version of the Nuttall window mentioned before. By using self convolution technique recursively, the attenuation of side lobe levels are doubled each iteration. Note that self convolution of a cosine series window is also a cosine series. It means that the spurious levels of closed-form anti-aliasing signals can be arbitrarily suppressed using this self convolution technique. However, -120 dB suppression is practically enough for many applications.

Also, note that the spurious level of the volume velocity-based over-sampling is practically aliasing-free, at least perceptually when taking masking into account. Those levels are also comparable to the side lobe level of the Blackman window.

V. Evaluation of F0 Extractors

This anti-aliased source wave serves as a flexible test signal for F0 extractors. Figures 12, 11 and 10 show preliminary examples of such tests. Figure 10 shows the test signal. In waveform plot, all segments are synchronously overlaid by setting the vocal fold opening as the reference point. It illustrates the amount of the temporal modulation. The spectrogram also illustrates the amount of the frequency modulation. The test signal is generated by using a F0 trajectory with frequency modulation. The average F0 is set to 220 Hz (A3 in musical note). The modulation frequency is 17 Hz and the modulation depth measured in semitones is one. A sinusoidal modulation is applied in the logarithmic frequency domain. The duration of the test signal is 1 s and the segment from 0.1 s to 0.9 s is used.

The modulation frequency 17 Hz used here may seem strange. It is selected to illustrate the difference of temporal resolution and the fidelity of F0 trajectory tracking when F0

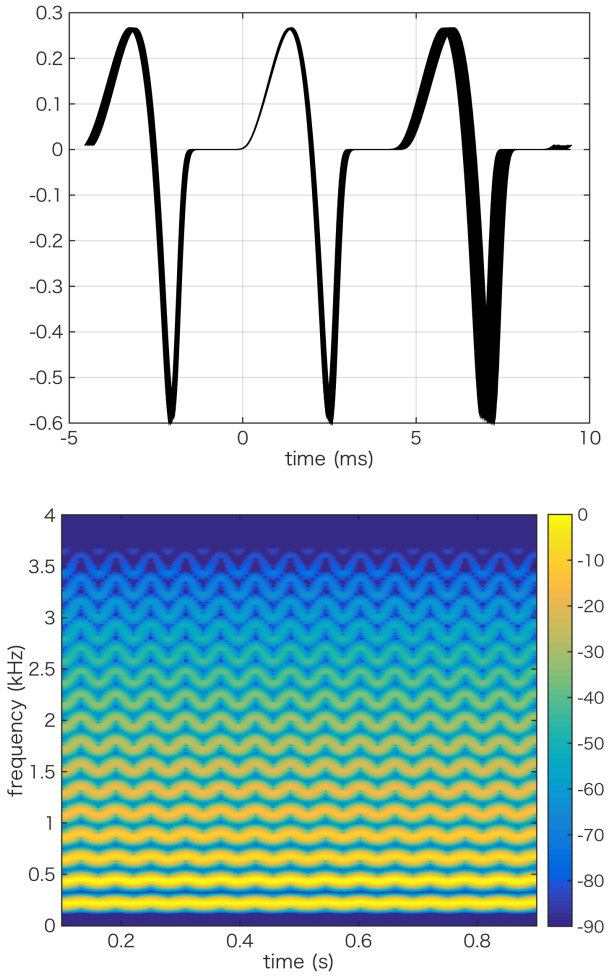

Fig. 10. Waveform and spectrogram of the generated test signal for F0 extractor evaluation. The waveform of each cycle is synchronously overlaid using the vocal fold opening as the time reference.

is modulated rapidly. Such rapid F0 modulations are found in extreme voices such as Noh, a Japanese classical theatrical performance and growl-like singing style in POP songs [21], [22]. For example, modulation frequency 70 Hz is found in a POP singing performance [23]. However, 70 Hz modulation cannot be handled by conventional F0 extractors properly. Consequently, 17 Hz is selected as a practical compromise. For sustained vowels, a comprehensive test is already reported [24] and there is no need for redundant tests.

This test signal is analyzed by several representative F0 extractors. The tested F0 extractors are YIN [25], SWIPEP [26], XSX [27], [28], higher-symmetry based method [23], [22] which consists of F0 refinement post-processing based on a interference-free representation of instantaneous frequency [29], and NDF [30], which is the optional F0 extractor of so-called legacy-STRAIGHT [31]. They are used in their default setting other than frame rate. The frame rate is set to 1 ms for all extractors. The time axis of YIN is adjusted by taking the buffer length into account. Figure. 11 shows the

```
%--- YIN ---
P.hop = 8;P.sr = fs;
R = yin(x,P);
f0Yin = 440*2.0.^(R.f0);
ttYin = (0:length(f0Yin)-1)'*R.hop/fs+R.wsize/fs/2;
%--- SWIPEP ---
[p,t,s] = swipep(x,fs);
%--- XSX --- default TANDEM-STRAIGHT
opt.framePeriod = 1;
r = exF0candidatesTSTRAIGHTGB(x,fs,opt);
%--- ICASSP 2013 and MAVEBA 2013 ---
f0Struct = higherSymKalmanWithTIFupdate(x,fs);
%--- NDF --- optional F0 for legacy-STRAIGHT
[f0raw,vuv,auxouts,prm]=MulticueF0v14(x,fs);
ttNDF = (0:length(f0raw)-1)/1000;
```

Fig. 11. Matlab script for analyzing the test signal by representative F0 extractors. The frame rate is set to 1 ms for all extractors.

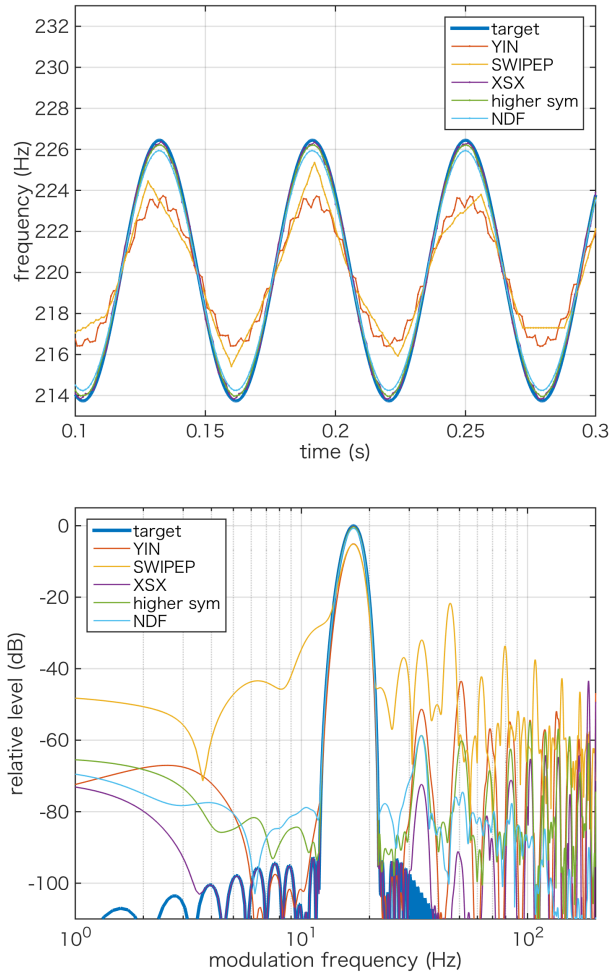

Fig. 12. The target and the extracted F0 trajectories by typical F0 extractors and their modulation frequency power spectrum.

Matlab script used in this set of analyses.

Figure 12 shows the test results. The upper plot of Fig. 12 shows the target and the extracted F0 trajectories. Note that the trajectories of the last three extractors are closely overlapping to the target trajectory. The F0 trajectories by YIN and SWIPEP look distorted, modulated and noisy. These observations are substantiated by modulation spectrum analysis.

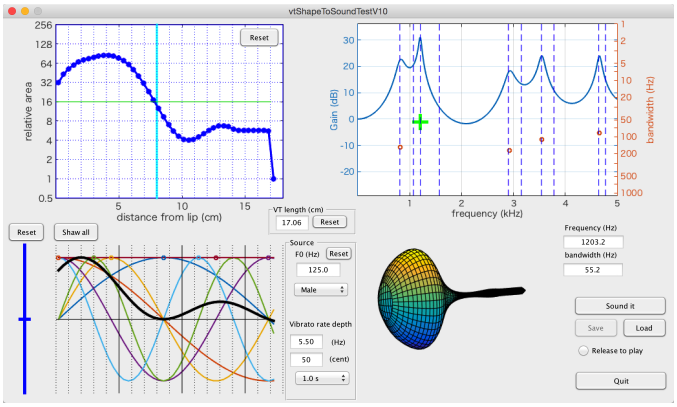

Fig. 13. A typical screen shot of the graphical user interface (GUI) of an interactive tool for investigating relations between vocal tract shape (including its length), transfer function of the corresponding one dimensional acoustic tube model (spectral shape and pole locations; frequencies and band widths), LSP frequencies and source information (F0, duration, vibrato rate and depth). (This shows OSX version)

The lower plot shows the modulation spectra of each trajectory. Prior to this analysis, F0 values are converted into logarithmic frequencies and Nuttall window is used for spectrum analysis. This modulation spectral plot indicates that the modulation spectral component other than the true modulation signal of the last three method are lower than -60 dB to the main component. This indicates that the last three methods, which are used in STRAIGHT systems, are precise trajectory trackers for time-varying F0 trajectories.

Note that these are preliminary results prepared to illustrate feasibility of F0 extractor evaluation based on this anti-aliased source signal representations. A systematic set of test for F0 extractors using this proposed signal is underway.

VI. Application to Speech Synthesis

When using this anti-aliased source signal in speech synthesis applications, additional spectral shaping is necessary. It is because the transfer function of the anti-aliasing filter based on Nuttall window introduces significant amount of higher frequency attenuation. This excessive attenuation has to be equalized. For example, for telephone band speech applications, spectral attenuation has to be close to 0 dB up to 3400 Hz. Such equalizer can be easily implemented using FIR filter design procedures based on windowing functions [32].

In this section, two possible examples are introduce to illustrate how the proposed signal can be useful for speech synthesis related systems. One is an interactive tool for speech science education. The other is STRAIGHT-based speech analysis, modification and synthesis systems [33].

A. *Interactive tool*

Figure 13 shows a typical screen shot of an educational tool for speech science. The anti-aliased source signal is to be integrated into this tool. The tool is designed to provide interactive environment to investigate relations between vocal tract shape

(including its length), transfer function of the corresponding one dimensional acoustic tube model [34], [35] (spectral shape and pole locations; frequencies and band widths), LSP (line spectrum pair) frequencies [36] (and underlying relations between LPC (linear prediction coefficients) family representations [37], [38], [39], [40]) and source information (F0, duration, vibrato rate and depth). This tool is accessible from the first author's web page with other Matlab realtime speech tools [41].

This investigation of the closed-form anti-aliased L-F model is motivated by this tool development. It was observed when high F0 frequency is set in this GUI, sometimes synthesized sounds sounded noisy and coarse. The anti-aliased version of the discrete L-F model will solve this problem and will provide additional control of voice quality by L-F model parameters.

B. STRAIGHT-based modification

It is well known that scaling law of formant frequencies and fundamental frequencies are different [42]. In current STRAIGHT-based manipulation, scaling of spectral shape and fundamental frequency are controlled differently. However, the spectral shape which STRAIGHT extracts consists of the vocal tract information and the glottal source information (so-called glottal formant, for example). Separation of the glottal source information from the STRAIGHT spectrum followed by manipulation of vocal tract related information and glottal source related information based on different scaling law is expected to improve re-synthesized speech quality. This is another prospective application of the proposed source signal.

VII. Discussion

There can be a possibility that the spectral equalization introduces perceptual artifacts, when taking into the nonlinear and onset sensitive nature of our auditory system [43], [44] into account. It is caused by so-called pre-echo [45]. Usual linear phase FIR filter is temporally symmetric and has preceding response before stimulation. The equalizer designed above inevitably consists of periodic oscillation around the corner frequency (for the telephone band case, 3400 Hz). It should be carefully tested two alternative implementation methods of this equalizer, linear phase FIR filter or minimum phase filter, which has no preceding response by definintion [32]. These are topics for further research.

VIII. Conclusions

A closed-form representation of anti-aliased L-F model is derived for a LPF function family based on cosine series. The Matlab based implementation of the derived form provides virtually aliasing-free source signal, which is applicable to speech synthesis and F0 extractor evaluation. This aliasing-free representation is also suitable for testing perceptual effects of wave shape parameters in the L-F model, since possible artifacts caused by spurious component are completely removed. A post processing procedure for fine tuning spectral shape is also introduced. An interactive tool for investigating speech production model parameters is designed using this Matlab implementation . This tool and the Matlab implementation of the proposed anti-aliased L-F model (with spectral equalization), together with tools for speech science education [41], are available from the first author's web page as an open source package [46].

Acknowledgment

This research is partly supported by Kakenhi (Aids for Scientific Research) of JSPS 15H02726, 15H03207, 25280063, 26284062 and 24650085.

References

[1] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech.*, vol. 4, pp. 1–13, 1985.

[2] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech.*, vol. 2-3, pp. 121–156, 1995.

[3] D. E. Sommer, I. T. Tokuda, S. D. Peterson, K.-I. Sakakibara, H. Imagawa, A. Yamauchi, T. Nito, T. Yamasoba, and N. Tayama, "Estimation of inferior-superior vocal fold kinematics from high-speed stereo endoscopic data in vivo," *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3290–3300, 2014.

[4] D. H. Klatt and L. C. Klatt, *Analysis, synthesis, and perception of voice quality variations among female and male talkers.*, 1990, vol. 87, no. 2.

[5] P. Alku, "Glottal inverse filtering analysis of human voice production —A review of estimation and parameterization methods of the glottal excitation and their applications," *SADHANA*, vol. 36, no. 5, pp. 623–650, 2011.

[6] X. Favory, N. Obin, G. Degottex, and R. Axel, "The role of glottal source parameters for high-quality transformation of perceptual age," in *ICASSP 2015*, Brisbane, Austolaria, 2015, pp. 4894–4898.

[7] D. G. Childers and C. Ahn, "Modeling the glottal volume - velocity waveform for three voice types," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995.

[8] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[9] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[10] P. H. Milenkovic, "Voice source model for continuous control of pitch period," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1087–1096, 1993.

[11] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *ICASSP 1986*, Tokyo, 1986, pp. 1605–1608.

[12] ——, "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," in *ICASSP 1987*, 1987, pp. 637–640.

[13] J. Walker and P. Murphy, "A review of glottal waveform analysis," in *Progress in nonlinear speech processing*. Springer, 2007, pp. 1–21.

[14] G. Degottex, A. Roebel, and X. Rodet, "Phase Minimization for Glottal Model Estimation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[15] Y. Shue, G. Chen, and A. Alwan, "On the Interdependencies between Voice Quality, Glottal Gaps, and Voice-Source related Acoustic Measures," in *Interspeech 2010*, no. September, 2010, pp. 34–37.

[16] D. Slepian, "Prolate spheroidal wave functions, Fourier analysis, and uncertainty-V: The discrete case," *Bell System Technical Journal*, vol. 57, no. 5, pp. 1371–1430, 1978.

[17] D. Slepian and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty-I," *Bell System Technical Journal*, vol. 40, no. 1, pp. 43–63, 1961.

[18] T. Stilson and J. Smith, "Alias-free digital synthesis of classic analog waveforms," in *Proc. International Computer Music Conference*, 1996, pp. 332–335.

[19] 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.

[20] 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.
[21] 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.
[22] H. Kawahara, M. Morise, and K. I. Sakakibara, "Temporally fine F0 extractor applied for frequency modulation power spectral analysis of singing voices," *8th International workshop: MAVEBA*, pp. 125–128, 2013. [Online]. Available: http://digital.casalini.it/an/2908956
[23] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Higher order waveform symmetry measure and its application to periodicity detectors for speech and singing with fine temporal resolution," in *ICASSP 2013*, 2013, pp. 6797–6801.
[24] A. Tsanas, M. Zañartu, M. a. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," *The Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 2885–2901, 2014.
[25] A. de Chevengné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.
[26] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.
[27] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," in *ICASSP 2008*, 2008, pp. 3933–3936.
[28] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT , a speech analysis , modification and synthesis framework," *SADHANA*, vol. 36, no. October, pp. 713–727, 2011.
[29] H. Kawahara, T. Irino, and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," in *ICASSP 2011*, May 2011, pp. 5420–5423.
[30] 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 2005*, 2005, pp. 537–540.
[31] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "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.
[32] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*, 3rd ed. Pearson, 2014.
[33] H. Kawahara, M. Morise, Banno, and V. G. Skuk, "Temporally variable multi-aspect N-way morphing based on interference-free speech representations," in *ASPIPA ASC 2013*, 2013, p. 0S28.02.
[34] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," *4th International Congress on Acoustics*, p. G42, 1962.
[35] H. Wakita, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 3, pp. 281–285, Jun. 1979.
[36] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.
[37] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencie," *Electro. Comm. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970.
[38] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.
[39] S. Sagayama and F. Itakura, "Duality theory of composite sinusoidal modeling and linear prediction," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, vol. 11, Apr. 1986, pp. 1261–1264.
[40] ——, "Symmetry between linear predictive coding and composite sinusoidal modeling," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 85, no. 6, pp. 42–54, 2002.
[41] H. Kawahara, "Matlab realtime tools for speech and signal processing education," *APSIPA Newsletter Issue-9*, pp. 5–10, 2015.
[42] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 2088–2113, 1989.
[43] J. O. Pickles, *An introduction to the physiology of hearing*, 6th ed. Brill Academic Pub., 2008.
[44] B. C. J. Moore, *An introduction to the psychology of hearing*, 6th ed. Brill Academic Pub., 2013.
[45] P. Noll, "MPEG digital audio coding," *Signal Processing Magazine, IEEE*, vol. 14, no. 5, pp. 59–81, 1997.
[46] H. Kawahara, "Matlab realtime speech tools and voice production tools." [Online]. Available: http://www.wakayama-u.ac.jp/%7ekawahara/MatlabRealtimeSpeechTools/

APPENDIX-A: DERIVATION OF CLOSED-FORM

The following derivation uses the unit step function $u(t)$ to represent piecewise functions. The open phase signal defined by (1) and the closing phase signal defined by (2) are represented as follows. Note that coefficients E_0 and $-E_0/\beta t_a$ are discarded here for making derivation simple. Similarly, the time origin of the closing phase is shifted to t_e. They are recovered when combining open phase and closing phase to compose the final form $E_b(t)$.

$$\begin{aligned} E_o(t) &= (u(t) - u(t-t_e))e^{\alpha t}\sin\omega_g t \\ &= (u(t) - u(t-t_e))e^{\alpha t}\frac{1}{2j}(e^{j\omega_g t} - e^{-j\omega_g t}), \quad (12) \\ E_s(t) &= (u(t) - u(t-t_d))(e^{\beta t} - e^{\beta t_d}), \quad (13) \end{aligned}$$

where the constant t_d, which represents the length of the decelerating closing phase is defined by $t_d = t_c - t_e$. The symbol j represents the imaginary unit $\sqrt{-1}$.

Anti-aliasing using cosine series

Let start from the general form of the anti-aliasing LPF $h(t)$, which is introduced by (4).

$$\begin{aligned} h(t) &= \sum_{k=0}^{N} w_k \cos\left(\frac{\pi k t}{T_w}\right), \quad \text{for} - T_w < t < T_w \\ &= (u(t+T_w) - u(t-T_w)) \sum_{k=0}^{N} w_k \cos\left(\frac{\pi k t}{T_w}\right) \\ &= \Re\left[(u(t+T_w) - u(t-T_w)) \sum_{k=0}^{N} w_k \exp\left(j\frac{\pi k t}{T_w}\right)\right], \quad (14) \end{aligned}$$

where T_w represents the half length of the support of the windowing function $h(t)$ and w_k represents the coefficient of the k-th harmonic cosine. It is convenient to represent $h(t)$ as the real part of the complex valued function $h_c(t)$, which is represented by the term located inside of the operator $\Re[\ ]$.

Using this complex valued anti-aliasing function $h_c(t)$ the anti-aliased version $x_b(t)$ of a real valued signal $x(t)$ is represented by the following equation.

$$x_b(t) = \Re\left[\int_{-\infty}^{\infty} h_c(\tau)x(t-\tau)d\tau\right] \quad (15)$$

By substituting the signals (12) and (13) and the anti-aliasing function (14) to (15) derivation starts. Applying integration by part and using periodicity of constituent complex exponents in $h_c(t)$ finally yields the following representations.

Closing phase signal representation

Let $E_{sb}(t)$ represent the anti-aliased version of $E_s(t)$. The derivation yields the following.

$$E_{sb}(t) = e^{\beta t}\sum_{k=0}^{N} w_k I_4(t) - e^{\beta t_d}\left(w_0 I_2(t) + \sum_{k=1}^{N} w_k I_{3r}(t)\right), \tag{16}$$

where constituent integrals $I_2(t), I_3(t)$ and $I_4(t)$ are given below.

The general term for the anti-aliasing of the second term in (13) has the following form.

$$I(t) = \int_{-T_w}^{T_w} (u(t-\tau) - u(t-t_d-\tau)) e^{j\frac{k\pi}{T_w}\tau} d\tau. \tag{17}$$

For $k = 0$, the exponential $e^{j\frac{k\pi}{T_w}\tau} = 1$ and yields the following.

$$\begin{aligned} I_2(t) = &- r(t-T_w) + r(t+T_w) \\ &+ r(t-t_d-T_w) - r(t-t_d+T_w), \end{aligned} \tag{18}$$

where $r(t)$ is the ramp function, which is 0 for $t \le 0$ and t for $t > 0$.

For $k > 0$, by using integration by parts, it yields.

$$\begin{aligned} I_3(t) &= \int_{-T_w}^{T_w} (u(t-\tau) - u(t-t_d-\tau)) e^{j\frac{k\pi}{T_w}\tau} d\tau \\ &= \frac{1}{j\xi}(u(t-\tau) - u(t-t_d-\tau)) e^{j\frac{k\pi}{T_w}\tau} \\ &\quad + \frac{1}{j\frac{k\pi}{T_w}} \int_{-T_w}^{T_w} (\delta(t-\tau) - \delta(t-t_d-\tau)) e^{j\frac{k\pi}{T_w}\tau} d\tau \\ &= \frac{1}{j\frac{k\pi}{T_w}} \left[(u(t-\tau) - u(t-t_d-\tau)) e^{j\frac{k\pi}{T_w}\tau}\right]_{-T_w}^{T_w} \\ &\quad + \frac{1}{j\frac{k\pi}{T_w}} \left(e^{j\frac{k\pi}{T_w}t}\Big|_{t\in\Omega_1} - e^{j\frac{k\pi}{T_w}(t-t_d)}\Big|_{t-t_d\in\Omega_1}\right), \end{aligned} \tag{19}$$

where the symbol Ω_1 represents the interval $[-T_w, T_w]$ and the notation $f(x)|_{P(x)}$ represents that the function $f(x)$ is defined when the logical predicate $P(x)$ is true.

Since $e^{j\frac{k\pi}{T_w}\tau} = e^{jk\pi}$ for $\tau = T_w$, and always a real number 1 or -1 for $k \in Z$ (where Z represents the set of integer), the first term vanishes when taking the real part of $I_3(t)$. The output signal is the real part of this complex signal. The real part of the k-th term $I_{3r}(t) = \Re[I_3(t)]$ is given by the following.

$$I_{3r}(t) = \frac{T_w}{k\pi}\left(\sin\left(\frac{k\pi t}{T_w}\right)\Big|_{t\in\Omega_1} - \sin\left(\frac{k\pi(t-t_d)}{T_w}\right)\Big|_{t-t_d\in\Omega_1}\right). \tag{20}$$

Similar derivation is also applicable to the first term in (13). Applying integration by part and the relation $e^{jk\pi} = (-1)^k$, it reduces to the following form.

$$\begin{aligned} I_4(t) = &(-1)^k \Re\left[\frac{1}{-\beta + j\frac{k\pi}{T_w}}\right] \cdot \\ &\left[e^{-\beta T_w}\big|_{t-T_w\in\Omega_2} - e^{\beta T_w}\big|_{t+T_w\in\Omega_2}\right] \\ &+ \Re\left[\frac{1}{-\beta + j\frac{k\pi}{T_w}} \cdot \right. \\ &\left.\left[e^{(-\beta + j\frac{k\pi}{T_w})t}\big|_{t\in\Omega_1} - e^{(-\beta + j\frac{k\pi}{T_w})(t-t_d)}\big|_{t-t_d\in\Omega_1}\right]\right], \end{aligned} \tag{21}$$

where $\Omega_2 = [0, t_d]$.

Open phase signal representation

Let $E_{ob}(t)$ represent the anti-aliased version of $E_o(t)$. The derivation yields the following.

$$E_{ob}(t) = \sum_{k=0}^{N} w_k \frac{e^{\alpha t}}{2} \Im\left[e^{j\omega_g t} I_5(t) - e^{-j\omega_g t} I_6(t)\right], \tag{22}$$

where constituent integrals $I_5(t)$ and $I_6(t)$ are given below. The derivation also uses integration by part and the relation $e^{jk\pi} = (-1)^k$.

$$\begin{aligned} I_5(t) = &\frac{1}{-\alpha + j\frac{k\pi}{T_w} - j\omega_g} \cdot \left[(-1)^k \right. \\ &\left[e^{(-\alpha - j\omega_g)T_w}\big|_{t-T_w\in\Omega_3} - e^{-(-\alpha - j\omega_g)T_w}\big|_{t+T_w\in\Omega_3}\right] \\ &+ e^{(-\alpha - j\omega_g + j\frac{k\pi}{T_w})t}\big|_{t\in\Omega_1} \\ &\left. - e^{(-\alpha - j\omega_g + j\frac{k\pi}{T_w})(t-t_e)}\big|_{t-t_e\in\Omega_1}\right], \end{aligned} \tag{23}$$

where $\Omega_3 = [0, t_e]$.

The part $I_6(t)$ is given below by replacing $-\omega_g$ with $+\omega_g$ in $I_5(5)$.

$$\begin{aligned} I_6(t) = &\frac{1}{-\alpha + j\frac{k\pi}{T_w} + j\omega_g} \cdot \left[(-1)^k \right. \\ &\left[e^{(-\alpha + j\omega_g)T_w}\big|_{t-T_w\in\Omega_3} - e^{-(-\alpha + j\omega_g)T_w}\big|_{t+T_w\in\Omega_3}\right] \\ &+ e^{(-\alpha + j\omega_g + j\frac{k\pi}{T_w})t}\big|_{t\in\Omega_1} \\ &\left. - e^{(-\alpha + j\omega_g + j\frac{k\pi}{T_w})(t-t_e)}\big|_{t-t_e\in\Omega_1}\right]. \end{aligned} \tag{24}$$

APPENDIX-B: L-F MODEL VOLUME VELOCITY

The volume velocity signal $v(t)$ based on the L-F model is given by integrating the original definition in (1) and (2).

$$v(t) = \frac{E_0 e^{\alpha t}(\alpha \sin\omega_g t - \omega_g \cos\omega_g t)}{\alpha^2 + \omega_g^2} + C_v \qquad t < t_e \tag{25}$$

$$v(t) = \frac{-e^{-\beta(t-t_e)}}{\beta} - (t-t_e)e^{-\beta(t_c-t_e)} + C_d \quad t_e \le t \le t_c, \tag{26}$$

where C_v and C_d are integral constants. They are determined to satisfy the following boundary conditions.

$$v(0) = 0 \tag{27}$$

$$v(t_c) = 0 \tag{28}$$