LETTER

Error Evaluation of an F0-Adaptive Spectral Envelope Estimator in Robustness against the Additive Noise and F0 Error

Masanori MORISE[†a)], ***Member***

SUMMARY This paper describes an evaluation of a temporally stable spectral envelope estimator proposed in our past research. The past research demonstrated that the proposed algorithm can synthesize speech that is as natural as the input speech. This paper focuses on an objective comparison, in which the proposed algorithm is compared with two modern estimation algorithms in terms of estimation performance and temporal stability. The results show that the proposed algorithm is superior to the others in both aspects.

key words: *speech analysis, spectral envelope, F0-adaptive analysis, time-varying component*

1. Introduction

High-quality speech synthesis is a major goal in speech research, and synthesizers for not only speech but also singing have been proposed and developed for consumer use. Several types of speech synthesis technologies have been proposed such as PSOLA [1], sinusoidal model [2], and phase vocoder. Since it is useful for various kinds of speech synthesis and voice conversion systems, the Vocoder-based [3] speech synthesis is dealt with in this paper. It consists of speech decomposition algorithms for fundamental frequency (F0), spectral envelope, and aperiodic parameter. The spectral envelope is a particularly important parameter and so is the dominant technique used in speech recognition, synthesis, and perception research. Several algorithms for high-quality speech synthesis and voice conversion have been proposed [4]–[8], and these algorithms would be effective for implementing real-time voice conversion systems [9], [10].

The reliable spectral envelope is also important in statistical parametric speech synthesis [11], [12], and a spectral envelope estimator using a statistical approach [13] has been proposed to improve the sound quality. Since the sound quality depends on the estimation performance of the spectral envelope, many researchers used a modern algorithm, called STRAIGHT [4], as the most reliable algorithm for estimating the spectral envelope. An accurate spectral envelope estimator is still useful for not only statistical parametric speech synthesis but also voice conversion such as voice morphing [14].

We have proposed a new speech analysis, manipulation, and synthesis framework [5] and several algorithms for spectral envelope estimation. The projects, named STRAIGHT-Library[*] and WORLD[**] are available on the website. Furthermore, aperiodicity estimation [15] and excitation signal estimation [16] have been proposed to improve the sound quality of synthesized speech. Since the proposed aperiodicity estimator requires an accurate spectral envelope, the performance of the spectral envelope must be estimated. This study is intended to evaluate the latest version of the spectral envelope estimator, named *Cheap-Trick* [17], from two objective aspects. In this paper, the robustness of the additive noise is evaluated. Since the algorithm uses an F0-adaptive window, the robustness of the error of the estimated F0 is also evaluated.

2. The Algorithm

First, the latest algorithm, named CheapTrick, is introduced. The concept of the algorithm is to obtain an accurate and temporally stable spectral envelope. In general speech analysis, the waveform and its power spectrum depend on the temporal position for windowing even if the all glottal vibrations and the spectral envelope are temporally stable. This temporal dependence is defined as the time-varying component and is dealt with as the error that should be removed in this paper. CheapTrick can remove the time-varying component and obtain the temporally stable spectral envelope.

As shown in Fig. 1, CheapTrick consists of three steps: F0-adaptive windowing, smoothing of the power spectrum, and a liftering processing for smoothing and spectral recovery. Since the algorithm was described in detail in another paper [17], this paper explains its outline.

2.1 First Step: F0-Adaptive Windowing

An F0-adaptive Hanning window is used on the basis of pitch synchronous analysis [18]. In CheapTrick, the length of the window is set to three times the fundamental period (T_0). The overall power of a periodic signal $y(t)$ windowed by the window $w(t)$ is calculated as follows:

$$\int_0^{3T_0} (y(t)w(t))^2\, dt \;=\; 1.125 \int_0^{T_0} y^2(t)dt. \qquad (1)$$

This equation shows that the overall power of the periodic

Manuscript received January 15, 2015.
Manuscript revised March 9, 2015.
Manuscript publicized April 2, 2015.

[†]The author is with the Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu-shi, 400–8511 Japan.
a) E-mail: mmorise@yamanashi.ac.jp

DOI: 10.1587/transinf.2015EDL8015

[*] http://ml.cs.yamanashi.ac.jp/straight/english/

[**] http://ml.cs.yamanashi.ac.jp/world/english/

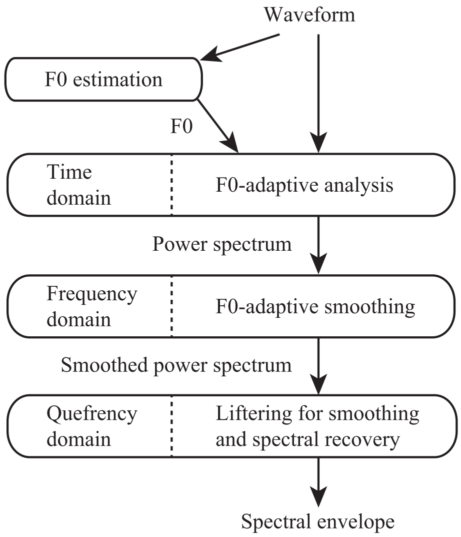

Fig. 1 Overview of the proposed algorithm.

signal windowed by the window is temporally stable.

2.2 Second Step: Frequency Domain Smoothing of the Power Spectrum

As shown in Fig. 1, the logarithmic power spectrum is used in the third step. The zero of the power spectrum indicates $-\infty$ in logarithmic power, which means that the zero in the power spectrum negatively influences the result of the third step. The spectral smoothing is defined as the preprocessing for the third step conducted to remove the zero in the power spectrum. This smoothing is given by

$$P_s(\omega) = \frac{3}{2\omega_0}\int_{-\frac{\omega_0}{3}}^{\frac{\omega_0}{3}} P(\omega+\lambda)d\lambda, \qquad (2)$$

where, $P(\omega)$ represents the power spectrum obtained in the first step.

2.3 Third Step: Liftering in the Quefrency Domain

In first and second steps, a power spectrum is obtained in which the overall power is temporally stable and has no zero. The third step is carried out to remove the time-varying component in all frequencies. Spectral recovery is required to compensate for the influence of the neighboring structure caused by the smoothing and is carried out on the basis of consistent sampling [19].

The final spectral envelope $P_l(\omega)$ is given by

$$P_l(\omega) = \exp\left(\mathcal{F}\left[l_s(\tau)l_q(\tau)p_s(\tau)\right]\right), \qquad (3)$$

$$l_s(\tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_0 \tau}, \qquad (4)$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1 \cos\left(\frac{2\pi\tau}{T_0}\right), \qquad (5)$$

$$p_s(\tau) = \mathcal{F}^{-1}\left[\log\left(P_s(\omega)\right)\right], \qquad (6)$$

where, $l_s(\tau)$ represents the liftering function for smoothing and $l_q(\tau)$ represents the liftering function for spectral recovery. $\tilde{q}_0$ and $\tilde{q}_1$ are the parameters for spectral recovery. Symbol $\mathcal{F}[]$ and $\mathcal{F}^{-1}[]$ represent Fourier transform and its inverse transform. In this paper, 1.18 and -0.09 are used as the values of $\tilde{q}_0$ and $\tilde{q}_1$ based on the prior work.

3. Evaluation

The results of subjective evaluation demonstrated that CheapTrick was superior to the conventional algorithms in terms of the sound quality of synthesized speech and robustness of the temporal F0 change [17]. On the other hand, since the real speech contains additive noise, the noise robustness must be verified. Since the exact F0 of real speech is difficult to estimate, this paper also evaluates CheapTrick in terms of the robustness of the F0 error.

3.1 Definition of the Evaluation Indexes

Two indexes in the estimation performance and the amount of time-varying component were employed for the evaluation. The index for evaluating the estimation performance is given by

$$E_f = \frac{1}{N}\sum_{n=0}^{N-1}\sigma_f(n), \qquad (7)$$

$$\sigma_f^2(n) = \frac{1}{K}\sum_{k=0}^{K-1}\left(P_e(k,n) - \frac{1}{K}\sum_{l=0}^{K-1}P_e(l,n)\right)^2, \qquad (8)$$

$$P_e(k,n) = 10\log_{10}\left(P_l(k,n)\right) - 10\log_{10}\left(P_t(k)\right), \qquad (9)$$

where $P_l(k,n)$ represents the time sequence of the estimated spectral envelope, k represents discrete frequency index, n represents the frame number, K represents the half value of FFT length, N represents the number of frames, and $P_t(k)$ represents the target spectral envelope.

The other index for evaluating the amount of time-varying component is given by

$$E_t = \frac{1}{K}\sum_{k=0}^{K-1}\sigma_t(k), \qquad (10)$$

$$\sigma_t^2(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left(P_e(k,n) - \frac{1}{N}\sum_{m=0}^{N-1}P_e(k,m)\right)^2. \qquad (11)$$

Both E_f and E_t indicate 0, provided that an accurate and temporally stable spectral envelope is obtained.

3.2 Experimental Conditions

To compare the performances of the proposed algorithm with those of other algorithms, the TANDEM-STRAIGHT and STAR [7] were used for the evaluation.

The target signal is a periodic impulse train with a standard F0. The target spectral envelope $P_t(k)$ is therefore flat in any temporal position for windowing. The sampling frequency of the signal is 48 kHz, FFT length is 4,096 samples,

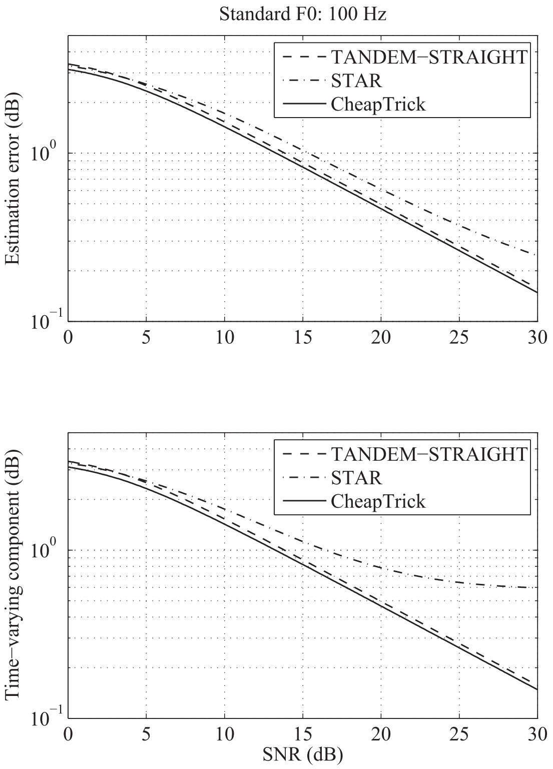

Fig. 2 Evaluation results in the robustness of the additive noise.

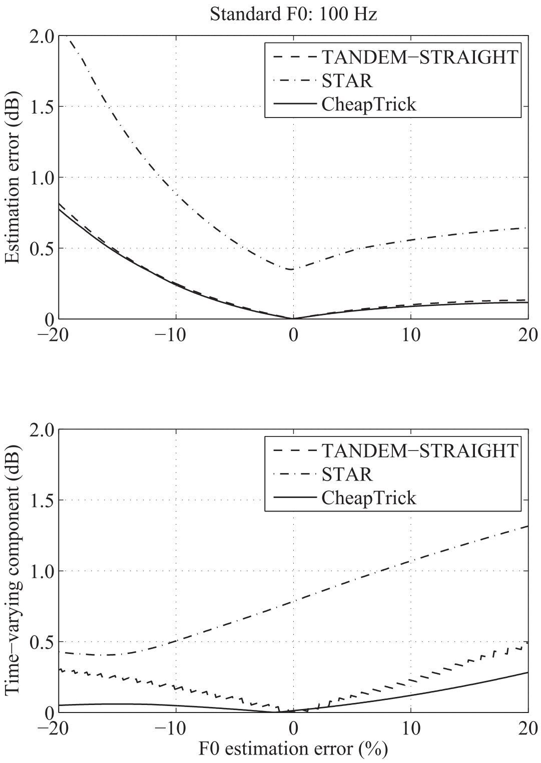

Fig. 3 Evaluation results in the F0 estimation error.

and the length of the signal is 1 s. The value of the frame shift is 1 ms, and the number of frames *N* is 1,000. To confirm the difference between the different F0, standard F0s of 100, 200, and 400 Hz were used for the evaluation. Since the results showed the same tendency, only the result of 100 Hz is illustrated and discussed in this paper.

3.3 Experiment 1: Robustness of the Additive Noise

In the experiment for verifying the robustness of the additive noise, SNR from 0 to 30 dB was used. A Gaussian white noise was used as the additive noise. Since the target spectral envelope was flat, the SNR was the same in all frequency bands. Figure 2 illustrates the results in the robustness of the additive noise. The horizontal and vertical axes represent the SNR and the evaluation indexes, respectively. The top and bottom represent the results for the estimation performance and the time-varying component, respectively.

The results of STAR were clearly inferior to those of other algorithms. CheapTrick was slightly superior to the TANDEM-STRAIGHT in both indexes regardless of SNR.

3.4 Experiment 2: Robustness of the F0 Error

In the evaluation in the F0 error, the error from −20 to 20% was used. Figure 3 illustrates the results in the robustness of the F0 error.

In the estimation performance, there were few differences between TANDEM-STRAIGHT and CheapTrick. On the other hand, CheapTrick was superior to other algorithms in the time-varying component, provided that the error was above ±3 dB. The results of TANDEM-STRAIGHT for the time-varying component seem to include a stepped variation. We confirmed that the variation was caused at the timings that the window length in sample points changed. The cause was the implementation of window function design for realizing the minor difference less than 1 sample.

3.5 Discussion

These results showed that CheapTrick was superior to conventional algorithms. The subjective evaluation has already been carried out, and its results suggested that CheapTrick achieved the best performance [17]. The results in this paper corroborated that CheapTrick was the best algorithm. In particular, since the major difference was the time-varying component, temporal stability may be the most dominant factor in high-quality speech synthesis.

4. Concluding Remarks

This paper assessed the temporally stable spectral envelope estimator, named CheapTrick, and its effectiveness was verified from two objective aspects. First, the details of the algorithm were explained, and then two evaluations were carried out to show its effectiveness. The results clearly showed that CheapTrick can robustly estimate the spectral envelope against F0 error and additive noise.

The next step is to apply CheapTrick to voice conversion and statistical parametric speech synthesis. Since we have proposed an aperiodicity estimator using the accurate spectral envelope [15], evaluation of the aperiodicity estimator is also important future work.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 24300073 and 26540087 and the Research Institute of Electrical Communication, Tohoku University (H25/A08).

References

[1] 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.

[2] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoustics, Speech, and Signal Processing, vol.34, no.4, pp.744–755, 1986.

[3] H. Dudley, "Remaking speech," J. Acoust. Soc. Am., vol.11, no.2, pp.169–177, 1939.

[4] 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.

[5] 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 Proc. ICASSP 2008, pp.3933–3936, 2008.

[6] H. Kawahara and M. Morise, "Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework," SADHANA - Academy Proceedings in Engineering Sciences, vol.36, no.5, pp.713–727, 2011.

[7] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," in Proc. SMAC 2013, pp.287–292, 2013.

[8] T. Nakano and M. Goto, "A spectral envelope estimation method based on f0-adaptive multi-frame integration analysis," in Proc. SAPA-SCALE 2012, pp.11–16, 2012.

[9] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime straight speech manipulation system: Report on its first implementation," Acoust. Sci. & Tech., vol.28, no.3, pp.140–146, 2007.

[10] M. Morise, M. Onishi, H. Kawahara, and H. Katayose, "v.morish'09: A morphing-based singing design interface for vocal melodies," Lecture Notes in Computer Science, vol.LNCS 5709, pp.185–190, 2009.

[11] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," Speech Communication, vol.51, no.11, pp.1039–1064, 2009.

[12] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proc. ICASSP 2013, pp.7962–7966, 2013.

[13] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory hmm," in Proc. ICASSP 2008, pp.3925–3928, 2008.

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

[15] H. Kawahara, M. Morise, H. Banno, R. Nisimura, and T. Irino, "Excitation source analysis for high-quality speech manipulation systems based on an interference-free representation of group delay with minimum phase response compensation," in Proc. INTERSPEECH 2014, pp.2243–2247, 2014.

[16] M. Morise, "Platinum: A method to extract excitation signals for voice synthesis system," Acoust. Sci. & Tech., vol.33, no.2, pp.123–125, 2012.

[17] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," Speech Communication, vol.67, pp.1–7, 2015.

[18] M.V. Mathews, J.E. Miller, and E.E. David, "Pitch synchronous analysis of voiced sounds," J. Acoust. Soc. Am., vol.33, pp.179–186, 1961.

[19] M. Unser, "Sampling — 50 years after shannon," Proc. of the IEEE, vol.88, pp.569–587, 2000.