CheapTrick, a spectral envelope estimator for high-quality speech synthesis

Article *in* Speech Communication · January 2014

DOI: 10.1016/j.specom.2014.09.003

1 author:

Masanori Morise
University of Yamanashi
111 PUBLICATIONS **499** CITATIONS

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 67 (2015) 1–7

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

CheapTrick, a spectral envelope estimator for high-quality speech synthesis

Masanori Morise

Interdisciplinary Graduate School of Medicine and Engineering, Univeristy of Yamanashi, 4-3-11, Takeda, Kofu, Yamanashi 400-8511, Japan

Received 14 May 2014; received in revised form 27 August 2014; accepted 9 September 2014
Available online 20 September 2014

Abstract

A spectral envelope estimation algorithm is presented to achieve high-quality speech synthesis. The concept of the algorithm is to obtain an accurate and temporally stable spectral envelope. The algorithm uses fundamental frequency (F0) and consists of F0-adaptive windowing, smoothing of the power spectrum, and spectral recovery in the quefrency domain. Objective and subjective evaluations were carried out to demonstrate the effectiveness of the proposed algorithm. Results of both evaluations indicated that the proposed algorithm can obtain a temporally stable spectral envelope and synthesize speech with higher sound quality than speech synthesized with other algorithms.

Keywords: Speech synthesis; Speech analysis; Spectral envelope; Time-varying component

1. Introduction

A speech analysis, manipulation, and synthesis framework is an important research topic ranging from singing synthesis (Kenmochi, 2012) to voice conversion. The origin of this framework was based on the idea of Vocoder (Dudley, 1939), and its framework consists of the fundamental frequency (F0) and spectral envelope estimators. In particular, since spectral envelope estimation was important for speech synthesis, several estimation algorithms such as Cepstrum (Noll, 1964; Oppenheim, 1969) and linear predictive coding (LPC) (Atal and Hanauer, 1969) have been proposed. Many improved algorithms have been proposed to estimate spectral envelopes accurately, such as discrete all-pole modeling (El-Jaroudi and Makhoul, 1991), weighted maximum likelihood autoregressive and moving average modeling (Badeau and David, 2009), and penalized likelihood approach (Campedel-Oudot et al., 2001). However, the sound quality of the speech synthesized by using these algorithms was less than that by using waveform-based synthesis such as PSOLA (Moulines and Charpentier, 1990). Sinusoidal models (McAulay and Quatieri, 1986) are also effective for high-quality speech synthesis, but it is difficult for them to achieve high-quality voice conversion. Spectral envelope is still useful for the flexible voice conversion such as eigenvoice conversion (Ohtani et al., 2010) and voice morphing (Kawahara et al., 2009).

Statistical parametric speech synthesis (Zen et al., 2009) is focused on for achieving a high-quality text-to-speech synthesis system. Several approaches have been proposed such as deep neural networks (Zen et al., 2013) and Gaussian process regression (Koriyama et al., 2014), and they require the accurate spectral envelope as the training data. Alternatively, a spectral envelope estimation method using a statistical approach (Toda and Tokuda, 2008) was also proposed for the statistical parametric speech synthesis.

A modern framework STRAIGHT (Kawahara et al., 1999) was proposed as an effective framework for high-quality speech synthesis. STRAIGHT achieved not only the high-quality speech synthesis but also F0 and spectral

E-mail address: mmorise@yamanashi.ac.jp

http://dx.doi.org/10.1016/j.specom.2014.09.003
0167-6393/

envelope manipulation without deterioration and has been used for voice conversion and singing synthesis. In particular, the spectral envelope estimated by STRAIGHT has been widely used for the statistical parametric speech synthesis. To reduce the computational cost while maintaining the sound quality, we proposed TANDEM-STRAIGHT (Kawahara et al., 2008; Kawahara and Morise, 2011) as an improved version of STRAIGHT. TANDEM-STRAIGHT made it possible to achieve the real-time application (Morise et al., 2009) and statistical analysis with a huge database.

High-quality speech synthesis has recently been focused on, and several spectral envelope estimation algorithms have been proposed (Morise, 2013; Nakano and Goto, 2012) for high-quality speech synthesis. In this paper, a simple algorithm for high-quality speech synthesis is introduced that is superior to conventional ones both objectively and subjectively.

The rest of this paper is organized as follows. In Section 2, we describe the problems facing the spectral envelope estimation. In Section 3, we propose a spectral envelope estimation method, named *CheapTrick*. In Section 4, we evaluate the proposed algorithm in objective and subjective evaluations. We conclude in Section 5 with a brief summary and a mention of future work.

2. Problems to be solved

According to the idea of STRAIGHT, which is an improvement of Vocoder, the voiced sound contains the F0, spectral envelope, and aperiodicity information. To simplify the discussion, we only deal with the fixed F0 and spectral envelope and leave out aperiodicity. The requirement is to obtain an accurate spectral envelope from speech waveform regardless of the temporal position for windowing.

Voiced sound $y(t)$ is approximated to the convolution of an impulse response $h(t)$ and periodic impulses. The spectrum of the waveform $Y(\omega)$ is given by

$$y(t) = h(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT_0), \tag{1}$$

$$Y(\omega) = \frac{2\pi}{T_0} H(\omega) \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0), \tag{2}$$

where, symbol * represents the convolution, T_0 represents the fundamental period, and ω_0 represents fundamental angular frequency ($= 2\pi/T_0$). $H(\omega)$ is the spectral envelope that is the estimation target. This equation shows that the target is discretized by convolving the periodic impulses in the time domain. On the other hand, Eq. (2) also shows that powers at $n\omega_0$ Hz (n: integer value) have exact values, which a spectral envelope estimation algorithm should therefore obtain. The spectral recovery is also required for compensating for the influence of discretization. Since the power spectrum of windowed waveform depends on the temporal position of the window function, this influence (time-varying component in this paper) should also be removed.

Conventional algorithms other than STRAIGHT and TANDEM-STRAIGHT cannot fulfill two requirements in the estimation performance and remove the time-varying component. This paper introduces an algorithm named *CheapTrick* that fulfills these requirements and is superior to conventional algorithms both objectively and subjectively. The name CheapTrick comes from its cheap and tricky design based on the conventional algorithms such as F0-adaptive windowing and the cepstrum method.

3. CheapTrick: detailed algorithm

CheapTrick consists of three steps: F0-adaptive windowing, smoothing of the power spectrum, and a liftering processing for smoothing and spectral recovery.

3.1. First step: F0-adaptive windowing

The first step is designing a window function on the basis of the ideal of pitch synchronous analysis (Mathews et al., 1961). A Hanning window with a length of $3T_0$ is used in the proposed algorithm. The power of the window at ω_0 Hz is 30 dB lower than that of the main lobe (0 Hz), which suggests that a harmonic structure influences the neighboring structure at below 30 dB. Since the actual speech contains temporal fluctuation, aperiodicity, and noise, we assumed that 30 dB was small enough.

The overall power of a periodic signal $y(t)$ windowed by the Hanning window $w(t)$ is calculated as follows;

$$\begin{aligned}\int_0^{3T_0} (y(t)w(t))^2 dt &= \int_0^{T_0} y^2(t)w^2(t)dt + \int_0^{T_0} y^2(t)w^2(t+T_0)dt \\ &\quad + \int_0^{T_0} y^2(t)w^2(t+2T_0)dt, \\ &= \int_0^{T_0} y^2(t)\big(w^2(t) + w^2(t+T_0) \\ &\quad + w^2(t+2T_0)\big)dt, \\ &= 1.125\int_0^{T_0} y^2(t)dt, \end{aligned} \tag{3}$$

where T_0 represents the period of the signal $y(t)$. This equation shows that the overall power of the periodic signal windowed by the window is temporally stable.

3.2. Second step: frequency domain smoothing of the power spectrum

The second step is the frequency domain smoothing of the power spectrum calculated in the first step. This step is positioned as the pre-processing of the third step carried out in the quefrency domain. A logarithmic power spectrum is used for the processing in the quefrency domain, but the zero of the power spectrum indicates $-\infty$ in logarithmic power and causes a fatal error. This step is carried

out to ensure that the power spectrum has no zeros. The smoothing is carried out by simple filtering with a rectangular window given by

$$P_s(\omega) = \frac{3}{2\omega_0}\int_{-\frac{\omega_0}{3}}^{\frac{\omega_0}{3}} P(\omega+\lambda)d\lambda, \tag{4}$$

where, $P(\omega)$ represents the power spectrum calculated from the waveform windowed by the Hanning window. Since the width of the rectangular window is $2\omega_0/3$, this step ensures that the influence between the neighboring structures is below 30 dB as in the first step.

3.3. Third step: liftering in the quefrency domain

The third step is liftering in the quefrency domain to remove the frequency fluctuation caused by discretization. The spectral recovery is also carried out at the same time. The time-varying component is caused by convolving the periodic impulses and windowing the waveform. Since the spectrum of periodic impulses with a period of T_0 is the periodic impulses with a period of ω_0, its cepstrum representation is also the periodic impulses with a period of T_0 in the quefrency domain. Fig. 1 shows an example in the time-varying component calculated in periodic impulses. The horizontal axis represents the normalized quefrency, and the vertical axis represents the standard deviation of the time series of cepstrum. To definitely calculate the time-varying component at nT_0, the sampling frequency of the signal is 65,536 Hz, F0 is 128 Hz, FFT length is 65,536, the length of the signal is 1 s, and the length of frame shift is one sample. Fig. 1 shows that the time-varying component converges at nT_0. To remove the time-varying component and to extract the low quefrency components, we employed a sinc function as the liftering that fulfills the requirements.

Spectral recovery is carried out on the basis of consistent sampling (Unser, 2000). Consistent sampling theory is the expansion of sampling theory and requires that the digital signal after AD conversion must equal the digital signal after AD/DA/AD conversion. This sampling enables the design of the linear filter to fulfill this requirement. In the proposed algorithm, the liftering function is designed as the compensation filter on the basis of consistent sampling. This function can compensate for the errors at $n\omega_0$ Hz, which are the sampling points caused by smoothing in both the second and third steps.

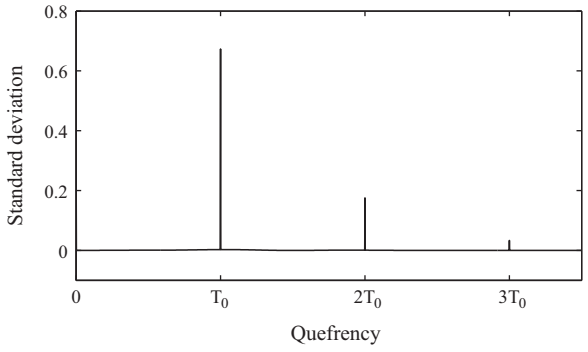

Fig. 1. A standard deviation of time series in the cepstrum of the periodic impulses. Time-varying component converges at nT_0 in the quefrency domain.

The spectral recovery based on consistent sampling has been carried out on the frequency domain in the TANDEM-STRAIGHT (Kawahara et al., 2008) as follows.

$$P_l(\omega) = \exp\left(\tilde{q}_0\log(P(\omega)) + \tilde{q}_1\log\left(P(\omega+\omega_0)P(\omega-\omega_0)\right)\right), \tag{5}$$

where $P_l(\omega)$ represents the output of spectral recovery processing, and $P(\omega)$ represents the power spectrum including the influence by smoothing. In this study, both the smoothing and spectral recovery were carried out in the quefrency domain. The final spectral envelope $P_l(\omega)$ is given by

$$P_l(\omega) = \exp\left(\mathcal{F}\left[l_s(\tau)l_q(\tau)p_s(\tau)\right]\right), \tag{6}$$

$$l_s(\tau) = \frac{\sin(\pi f_0\tau)}{\pi f_0\tau}, \tag{7}$$

$$l_q(\tau) = \tilde{q}_0 + 2\tilde{q}_1\cos\left(\frac{2\pi\tau}{T_0}\right), \tag{8}$$

$$p_s(\tau) = \mathcal{F}^{-1}[\log\left(P_s(\omega)\right)], \tag{9}$$

where, $l_s(\tau)$ represents the liftering function for smoothing and $l_q(\tau)$ represents the liftering function for spectral recovery derived from Eq. (5). $\tilde{q}_0$ and $\tilde{q}_1$ are the parameters for spectral recovery. Symbol $\mathcal{F}[]$ and $\mathcal{F}^{-1}[]$ represent Fourier transform and its inverse transform.

$l_s(\tau)$ shows that the smoothing is carried out by convolving the rectangular window on the frequency domain. This window has zeros at nT_0 in the quefrency domain shown in Fig. 2. To determine the two parameters in $l_q(\tau)$, an exploratory evaluation was carried out. The conditions were same as the following objective evaluation. As a result, 1.18 and −0.09 are obtained as the values of $\tilde{q}_0$ and $\tilde{q}_1$, respectively, and used in the following evaluation.

Each processing is carried out on discrete time–frequency representation. $P_l(k,n)$ discretized on both time and frequency domains is therefore used, where k represents discrete frequency index and n represents the frame number.

4. Evaluation and discussion

To verify the effectiveness of the proposed algorithm (CheapTrick), both objective and subjective evaluations were carried out. The TANDEM-STRAIGHT and STAR (Morise, 2013) were used for comparison. The objective evaluation used two evaluation indexes in the estimation accuracy and time-varying component. In the subjective evaluation, a MUSHRA-based evaluation and a Thurstone's paired comparison were carried out to demonstrate the effectiveness of the proposed algorithm in terms of sound quality of not only the re-synthesized but also the F0-manipulated speech.

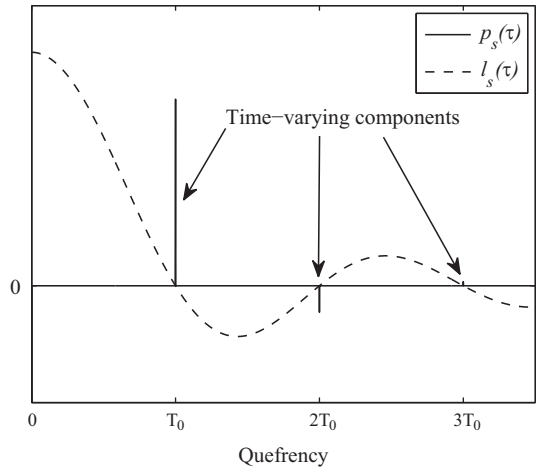

Fig. 2. The cepstrum $p_s(\tau)$ and the liftering function $l_s(\tau)$. Since the liftering function has zeros at nT_0, the time-varying component is removed.

4.1. Objective evaluation

4.1.1. Definition of the evaluation indexes

The index for evaluating the estimation accuracy is given by

$$E_f = \frac{1}{N}\sum_{n=0}^{N-1}\sigma_f(n), \tag{10}$$

$$\sigma_f^2(n) = \frac{1}{K}\sum_{k=0}^{K-1}\left(P_e(k,n) - \frac{1}{K}\sum_{l=0}^{K-1}P_e(l,n)\right)^2, \tag{11}$$

$$P_e(k,n) = 10\log_{10}(P_l(k,n)) - 10\log_{10}(P_t(k)), \tag{12}$$

where N represents the number of frames and K represents the half value of FFT length. $P_t(k)$ represents the target spectral envelope. E_f indicates 0, provided that the estimation result equals the target.

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

$$E_t = \frac{1}{K}\sum_{k=0}^{K-1}\sigma_t(k), \tag{13}$$

$$\sigma_t^2(k) = \frac{1}{N}\sum_{n=0}^{N-1}\left(P_e(k,n) - \frac{1}{N}\sum_{m=0}^{N-1}P_e(k,m)\right)^2. \tag{14}$$

E_t also indicates 0, provided that the time-varying component was completely removed. The algorithm that can minimize both E_f and E_t is the best one in this paper.

4.1.2. Experiment 1: robustness of the F0 change

In the experiment 1, the test signals were designed to confirm the robustness of the F0 change because the F0 of speech is time-varying. In this experiment, test signals were used that were designed using the following equation including the parameter for controlling the gradient.

$$f_0(t) = f_c + \sqrt{\alpha f_c}\cos\left(\sqrt{\alpha f_c}t + \theta\right), \tag{15}$$

where f_c represents the standard F0, θ represents the phase, and α represents the parameter for controlling the gradient. In this F0 contour, the maximum gradient is αf_c. We used the standard F0s of 100, 200, and 400 Hz to confirm the dependency of F0; α from 0.0 to 36.0 was used. θ from 0.0 to 2π was used to calculate evaluation indexes at a temporal position, and results calculated in all θ were averaged to obtain the definitive result.

Since the signal for experiment is an aperiodic impulse train, the target spectrum is flat without depending on the temporal position for windowing. The sampling frequency of the signal is 48 kHz, FFT length is 4096 samples, and the length of the signal is 1 s. The value of the frame shift is 1 ms, and the number of frames is 1000.

Figs. 3 and 4 illustrate the evaluation results of estimation accuracy and the time-varying component, respectively. The results show that the STAR was clearly inferior to the other algorithms in both evaluation indexes. In all results, the errors in higher F0 are smaller than those in lower F0. CheapTrick outperformed both other algorithms in estimation accuracy regardless of α and standard F0. From the results of the time-varying component, CheapTrick was the best in the standard F0 of 100 Hz. CheapTrick and TANDEM-STRAIGHT exhibited almost the same performance for other standard F0s.

4.1.3. Experiment 2: relationship between the spectral slope and the estimation indexes

In experiment 2, the test signals were designed to confirm the relationship between the estimation performance and the spectral slope. The target spectral envelope has spectral slope of $-\beta$ dB/oct from 100 Hz. β from 0.0 to 10.0 was used, and Eq. (15) with α of 10 was used to generate the F0 contour. α was set for comparing results under realistic conditions because F0 of real speech is time-varying. Since the same tendencies were observed in the experiment, a standard F0 of 100 Hz was employed. Other conditions were the same as those for experiment 1.

Fig. 5 illustrates the relationship between the parameter β and the evaluation indexes. In the time-varying component, CheapTrick was superior to other algorithms, and it was relatively better than TANDEM-STRAIGHT, provided that β was above 8.5. In the estimation error, all algorithms performed almost the same, provided that β was above 8.5. These results suggested that the proposed algorithm was superior for estimating the spectral envelope including sharp formants.

4.2. Subjective evaluation

Two subjective evaluations were carried out to verify the sound quality of not only the re-synthesized but also the F0-manipulated speech. Since the F0 and the aperiodicity estimated by TANDEM-STRAIGHT are used in the evaluation, the sound quality of the speech only depends on the estimated spectral envelope. The speech used for the evaluation was of three males and three females from

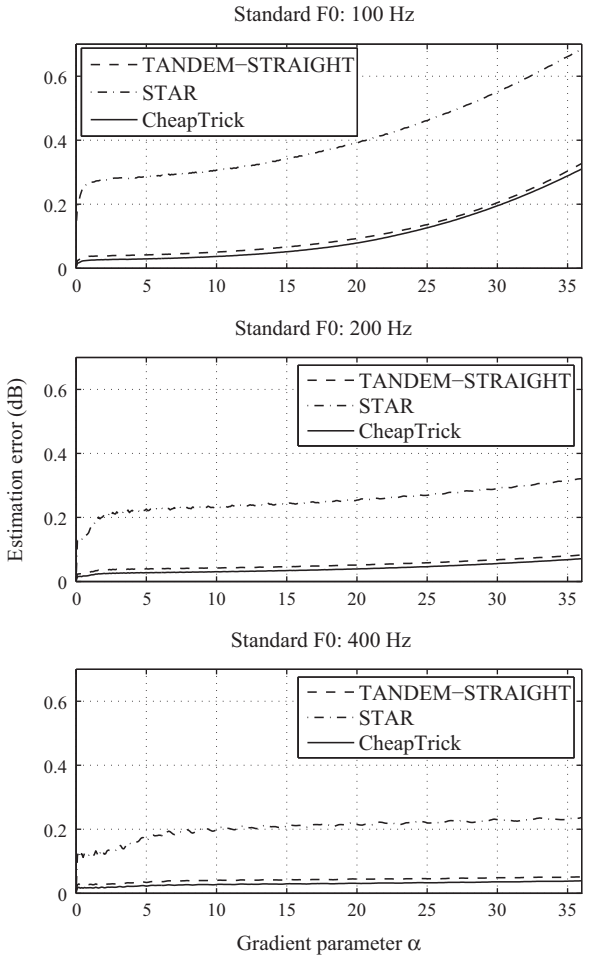

Fig. 3. Objective evaluation results of estimation accuracy.

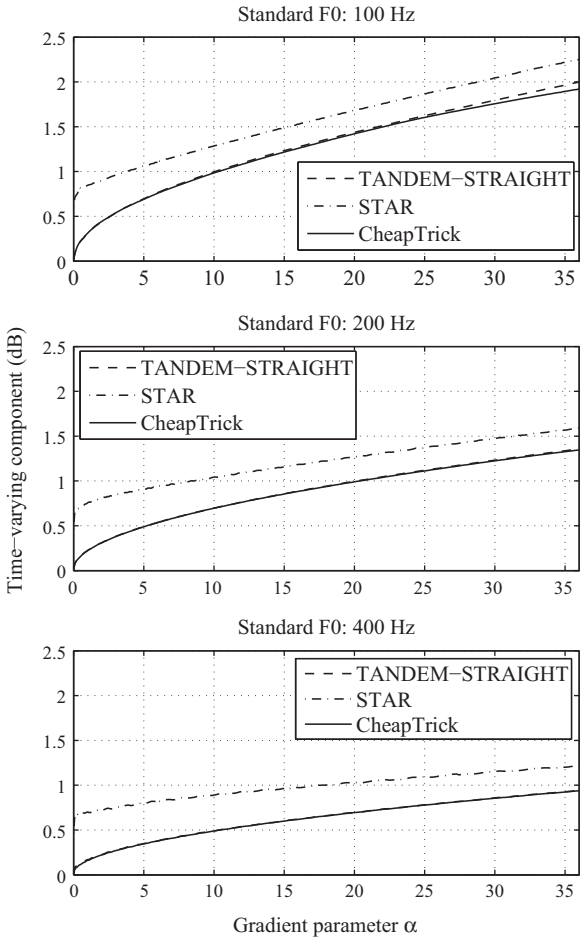

Fig. 4. Objective evaluation results of time-varying component.

a DVD of a Japanese-language textbook. The sampling was 44.1 kHz/16 bit, and an anechoic room was used for the recording. The speech was five Japanese vowels /aiueo/ continuously uttered by all speakers and did not include consonants because all algorithms were for the spectral envelope estimation of voiced speech. In the subjective evaluations, a sound proof room with the A-weighted SPL of 16-dB was used. Ten subjects with normal hearing ability participated in both evaluations.

A MUSHRA evaluation based on the ITU-R recommendation BS.1534-1 was carried out to compare the sound quality of original and re-synthesized speech. Since MUSHRA uses the 0–100 scale, subjects can rate smaller differences than when using MOS evaluation. The subjects evaluated all speech by using a GUI that displayed four kinds of stimuli (the original speech and speech synthesized with TANDEM-STRAIGHT, STAR, and CheapTrick) at the same time. Therefore, the subjects could evaluate speech uttered by the same speaker. When evaluating the four sets of speech by the GUI is defined as one set of stimuli, the number of sets is six.

A Thurstone's paired comparison was carried out by using not only the re-synthesized but also the F0-manipulated speech. In this evaluation, the subjects were instructed

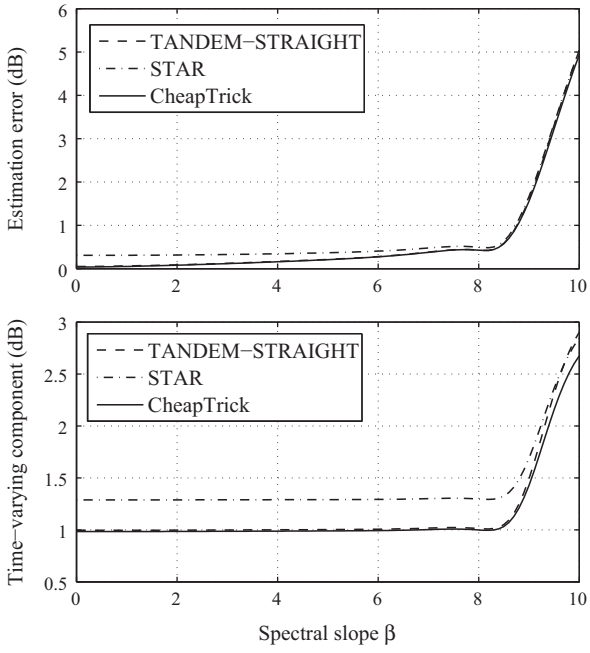

Fig. 5. Relationship between the spectral slope β and the evaluation indexes. The standard F0 was 100 Hz.

to select the stimulus with higher sound quality. Three types of F0 (F0, 1.25 F0, and 0.75 F0) were used to demonstrate the robustness against F0-manipulation. In the three sets of speech synthesized with the same F0, all combinations excluding the same stimulus pairs were used. The order effect is counterbalanced by using both A–B and B–A pairs. The stimulus pairs numbered 108, and all were randomized.

4.3. Results and discussion

Figs. 6 and 7 illustrate the experimental results. The MUSHRA evaluation results showed that three algorithms cannot synthesize speech as natural as the input speech, but each algorithm achieves more than 90. In female speech, the input speech did not significantly differ from the speech synthesized with CheapTrick. On the other hand, the input speech significantly differed from the speech synthesized with other algorithms. In all conditions, the three algorithms had no significant differences. This result showed that all algorithms can synthesize high-quality speech equally.

The results of Thurstone's paired comparison showed CheapTrick was superior to the other algorithms in terms of sound quality regardless of gender. Since the results include the sound quality of not only the re-synthesized speech but also the F0-manipulated speech, they suggested that CheapTrick was robust against F0 manipulation. The difference in sound quality in female speech was smaller than that in male speech, and this difference is associated with the objective evaluation results in which the error in higher F0 was smaller than that in lower F0.

Although we evaluated synthetic speech of continuously uttered vowels, the synthetic speech must be evaluated for not only the vowels but also the consonants. However, the proposed algorithm is for the voiced sound that has a period, and general vocoder processes the voiced and unvoiced sound separately. Since the processing for unvoiced sound affects the sound quality, we did not conduct the evaluation using speech including consonants to avoid the influence

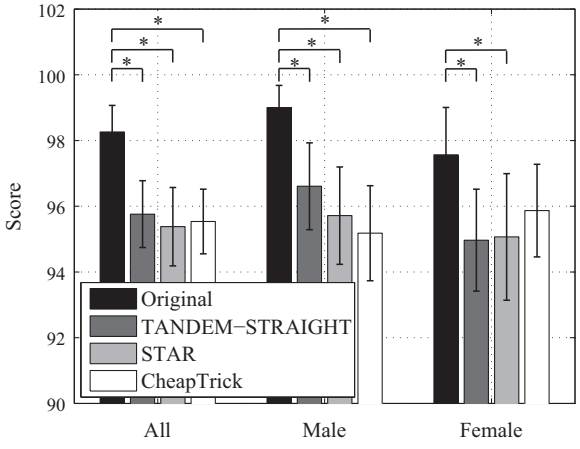

Fig. 6. Subjective evaluation results in (Left) evaluation (symbol * represents significant difference ($p < 0.05$), and error bars represent 95% confidence intervals).

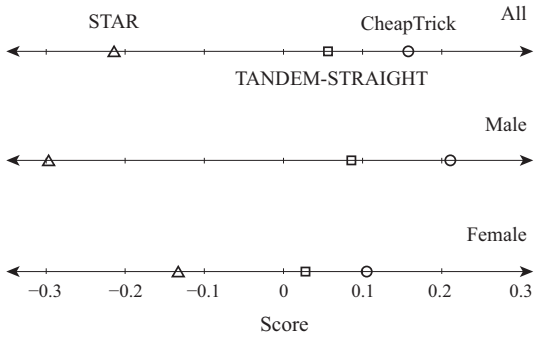

Fig. 7. Subjective evaluation results in Thurstone's paired comparison.

caused by any methods other than spectral envelope estimation. An algorithm needs to be developed to estimate the spectral envelope of consonants, and algorithms need to be evaluated with vowels and consonants.

5. Concluding remarks

A spectral envelope estimation algorithm was proposed for high-quality speech synthesis. This paper started by examining the well-known problem of the spectral envelope estimation and discussed the concept of the proposed algorithm named CheapTrick. CheapTrick consists of power spectrum estimation with the F0-adaptive Hanning window, the smoothing of the power spectrum, and spectral recovery in the quefrency domain. The algorithm can obtain an accurate and temporally stable spectral envelope by objective evaluations.

The subjective evaluations demonstrated that CheapTrick was superior to the conventional algorithms. In particular, CheapTrick synthesized the F0-manipulated speech more robustly than the other algorithms. Applying CheapTrick for voice morphing and statistical parametric speech synthesis is important future work.

Acknowledgements

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

References

Atal, B.S., Hanauer, S.L., 1969. Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Am. 50, 637–655.

Badeau, R., David, B., 2009. Weighted maximum likelihood autoregressive and moving average spectrum modeling. In: Proceedings of the ICASSP 2008, pp. 3761–3764.

Campedel-Oudot, M., Cappé, O., Moulines, E., 2001. Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach. IEEE Trans. Speech Audio Process. 9, 469–481.

Dudley, H., 1939. Remaking speech. J. Acoust. Soc. Am. 11, 169–177.

El-Jaroudi, A., Makhoul, J., 1991. Discrete all-pole modeling. IEEE Trans. Signal Process. 39, 411–423.

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing

and an instantaneous-frequency-based f0 extraction, speech communication. Speech Commun. 27, 187–207.
Kawahara, H., Morise, M., 2011. Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework. SADHANA – Acad. Proc. Eng. Sci. 36, 713–728.
Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H., 2008. Tandem-straight: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In: Proceedings of the ICASSP 2008, pp. 3933–3936.
Kawahara, H., Nisimura, R., Irino, T., Morise, M., Takahashi, T., Banno, H., 2009. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. In: Proceedings of ICASSP2009, pp. 3905–3908.
Kenmochi, H., 2012. Singing synthesis as a new musical instrument. In: Proceedings of the ICASSP 2012, pp. 5385–5388.
Koriyama, T., Nose, T., Kobayashi, T., 2014. Statistical parametric speech synthesis based on gaussian process regression. IEEE J. Select. Top. Sig. Process. 8, 173–183.
Mathews, M.V., Miller, J.E., David, E.E., 1961. Pitch synchronous analysis of voiced sounds. J. Acoust. Soc. Am. 33, 179–186.
McAulay, R., Quatieri, T., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust., Speech, Sig. Process. 34, 744–755.
Morise, M., 2013. An attempt to develop a singing synthesizer by collaborative creation. In: Proceedings of the SMAC 2013, pp. 287–292.
Morise, M., Onishi, M., Kawahara, H., Katayose, H., 2009. v.morish'09: a morphing-based singing design interface for vocal melodies. In: Lecture Notes in Computer Science LNCS, vol. 5709, pp. 185–190.
Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453–467.
Nakano, T., Goto, M., 2012. A spectral envelope estimation method based on f0-adaptive multi-frame integration analysis. in: Proceedings of the SAPA-SCALE 2012, pp. 11–16.
Noll, A.M., 1964. Short-time spectrum and "cepstrum" techniques for vocal pitch detection. J. Acoust. Soc. Am. 36, 296–302.
Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2010. Improvements of the one-to-many eigenvoice conversion system. IEICE Trans. Inform. Syst. E93-D, 2491–2499.
Oppenheim, A.V., 1969. Speech analysis-synthesis system based on homomorphic filtering. J. Acoust. Soc. Am. 45, 458–465.
Toda, T., Tokuda, K., 2008. Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM. In: Proceedings of the ICASSP 2008, pp. 3925–3928.
Unser, M., 2000. Sampling – 50 years after Shannon. Proc. IEEE 88, 569–587.
Zen, H., Senior, A., Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the ICASSP 2013, pp. 7962–7966.
Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51, 1039–1064.