IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2015EDP7457

Publicized:2016/04/05

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

WORLD: a vocoder-based high-quality speech synthesis system for real-time applications

Masanori MORISE[†], ***Member***, **Fumiya YOKOMORI**[††], ***Nonmember***, ***and*** **Kenji OZAWA**[†], ***Member***

SUMMARY A vocoder-based speech synthesis system, named WORLD, was developed in an effort to improve the sound quality of real-time applications using speech. Speech analysis, manipulation, and synthesis on the basis of vocoders are used in various kinds of speech research. Although several high-quality speech synthesis systems have been developed, real-time processing has been difficult with them because of their high computational costs. This new speech synthesis system has not only sound quality but also quick processing. It consists of three analysis algorithms and one synthesis algorithm proposed in our previous research. The effectiveness of the system was evaluated by comparing its output with against natural speech including consonants. Its processing speed was also compared with those of conventional systems. The results showed that WORLD was superior to the other systems in terms of both sound quality and processing speed. In particular, it was over ten times faster than the conventional systems, and the real time factor (RTF) indicated that it was fast enough for real-time processing.

key words: *speech analysis, speech synthesis, vocoder, sound quality, real-time processing*

1. Introduction

High-quality speech synthesis systems are being used in various applications, such as singing synthesizers [1] and voice conversion systems [2]. In particular, speech analysis, manipulation, and synthesis based on the idea of the vocoder [3] are widely used. Such systems consist of fundamental frequency (F0) and spectral envelope estimation algorithms and a synthesis algorithm that takes the estimated speech parameters. However, the speech synthesized by most of the conventional vocoder systems is inferior to that of waveform-based systems [4]. An exception is a vocoder-based system called STRAIGHT [5], which is capable of high-quality speech synthesis. STRAIGHT also makes it easy to manipulate speech, and its technology has been used in various studies. Other speech synthesis systems have been proposed such as phase vocoder [6], PSOLA [7], and sinusoidal model [8], and high-quality speech synthesis remains a popular research topic.

Real-time processing is another topic of speech synthesis research. For example, voice conversion for Karaoke [9] requires real-time analysis and synthesis. Real-time STRAIGHT [10] has been proposed as a way to meet the demand for real-time processing, but the simplified algorithm it uses degrades the quality of the synthesized speech. Real-time singing morphing [11] has the same problem. TANDEM-STRAIGHT [12], [13] is supposed to be a simplified version that outputs almost all the same parameters as STRAIGHT. Although the system works well, it is hard to use it for real-time speech analysis and synthesis.

This paper describes a high-quality speech synthesis system, named WORLD, to meet the requirements of not only high sound quality but also real-time processing. WORLD consists of three algorithms for obtaining three speech parameters and a synthesis algorithm that takes these parameters as input. Our previous research [14]–[17] carried out individual evaluations on WORLD. In this paper, we report a series of evaluations of its sound quality and processing speed. We also discuss its effectiveness on the basis of the results.

The rest of this paper is organized as follows. In Section 2, we give an overview of WORLD and the differences between it and conventional systems. In Section 3, we evaluate it in terms of sound quality and processing speed. In Section 4, we discuss its effectiveness. Finally in Section 5, we conclude with a brief summary and mention of future work.

2. Overview of WORLD

Figure 1 illustrates the processing of the system. First, the F0 contour is estimated with DIO [14], [15]. Second, the spectral envelope is estimated with CheapTrick [16], [18], which uses not only the waveform but also the F0 information. Third, the excitation signal is estimated with PLATINUM and used as an aperiodic parameter [17]. PLATINUM uses the waveform, F0, and spectral envelope information. Its definition of the aperiodic parameter is different from that of STRAIGHT (here, it is referred to as Legacy-STRAIGHT) or that of TANDEM-STRAIGHT.

WORLD cannot manipulate the aperiodic parameter as well as Legacy-STRAIGHT or TANDEM-STRAIGHT. On the other hand, it can manipulate the F0 and spectral envelope in the same manner as them [17]. The features of each algorithm are briefly explained below.

2.1 DIO: F0 estimation algorithm [14], [15]

F0 is defined as the inverse of the smallest period of a periodic signal. F0 estimation is a topic of speech analysis, and

[†]The authors are with the Interdisciplinary Graduate School, University of Yamanashi, 4-3-11 Takeda, Kofu 400-8511, Japan.

[††]The author is with the Graduate School of Medicine and Engineering Science Department of Education, University of Yamanashi, 4-3-11 Takeda, Kofu 400-8511, Japan.

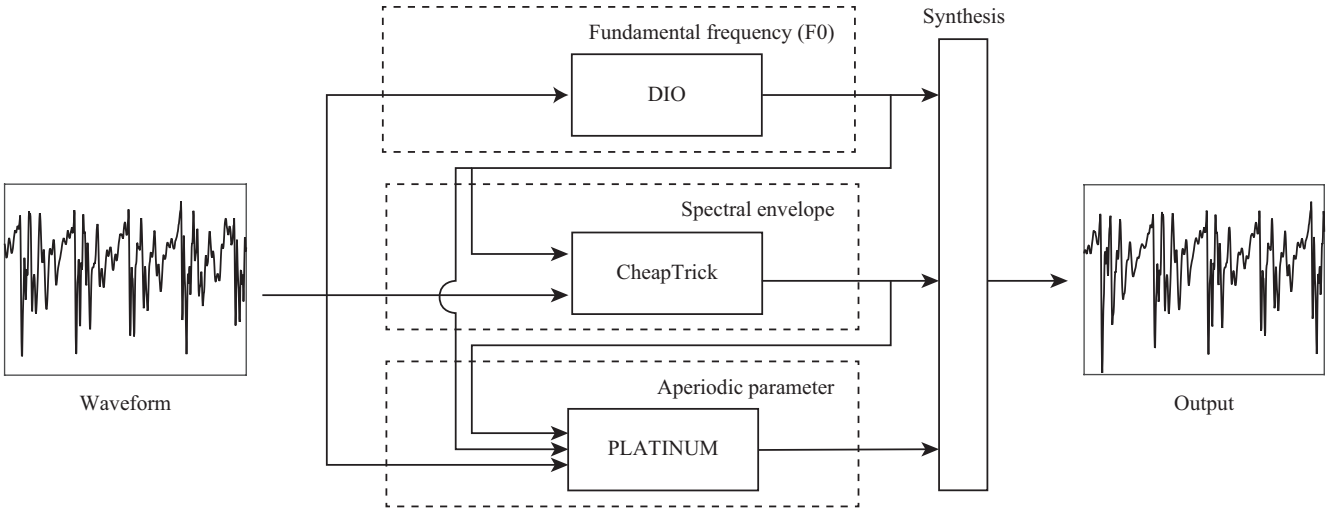


Fig. 1 Overview of the developed system. WORLD consists of three analysis algorithms for determining the F0, spectral envelope, and aperiodic parameters and a synthesis algorithm incorporating these parameters.

many algorithms have been proposed for it [19]. There are two particular types of characteristics: temporal characteristics, such as auto-correlation, and spectral characteristics such as Cepstrum [20].

WORLD uses a rapid and reliable F0 estimation algorithm, named DIO [14]. Even though it is much faster than other algorithms such as YIN [21] and SWIPE [22], its estimation performance is not inferior to theirs [15].

DIO consists of three steps. The first step is low-pass filtering with different cutoff frequencies. If the filtered signal only consists of the fundamental component, it forms a sine wave with a period of T_0, which is the fundamental period. Since the target F0 is unknown, many filters with different cutoff frequencies are used in this step.

The second step is to calculate the F0 candidates and their reliabilities in each filtered signal. Since a signal that consists of only the fundamental component forms a sine wave, the four intervals of the waveform, i.e., the positive and negative zero-crossing intervals and peak and dip intervals, shown in Fig. 2 have the same value. Their standard deviation is therefore associated with the reliability measure, and their average is defined as an F0 candidate. In the third step, the candidate with the highest reliability is selected.

2.2 CheapTrick: Spectral envelope estimation algorithm [16], [18]

The spectral envelope is also an important parameter in speech processing, and Cepstrum and linear predictive coding (LPC) [23] are typical algorithms for determining it. Although many algorithms based on these ones have been developed, they are not able to synthesize natural speech. The main problem is that the estimated result depends on the temporal position, so it is crucial to remove its influence, i.e., the time-varying component, while maintaining

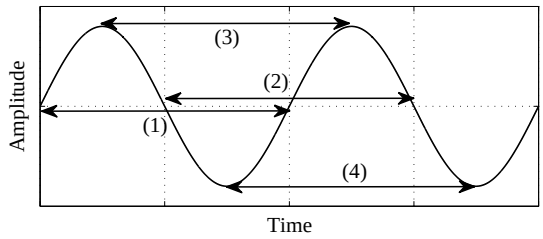


Fig. 2 Four intervals used for calculating an F0 candidate and its reliability. If the filtered signal only consists of the fundamental component, the four intervals indicate the same value.

estimation accuracy. Legacy-STRAIGHT and TANDEM-STRAIGHT were developed as algorithms to meet the requirements for high-quality speech synthesis. A number of other algorithms [24], [25] have been proposed for the same purpose.

WORLD uses an accurate spectral envelope estimation algorithm called CheapTrick [16], [18]. CheapTrick is based on the idea of pitch synchronous analysis [26] and uses a Hanning window with the length of $3T_0$. First, the power spectrum is calculated on the basis of the windowed waveform. The overall power of the windowed waveform is temporally stabilized using the following equation:

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

where $y(t)$ represents the waveform, and $w(t)$ represents the window function.

Then, the power spectrum is smoothed with a rectangular window of width $2\omega_0/3$, as follows:

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

where ω_0 is defined as $2\pi/T_0$.

Finally, specialized liftering is carried out:

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

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

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

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

where $l_s(\tau)$ represents the filtering function for smoothing the logarithmic power spectrum and removing the time-varying component. $l_q(\tau)$ represents the liftering function for spectral recovery. This recovery function improves the estimation performance. $p_s(\tau)$ represents the Cepstrum of the power spectrum $P_s(\omega)$. The symbols $\mathcal{F}[]$ and $\mathcal{F}^{-1}[]$ represent the Fourier transform and its inverse transform. $\tilde{q}_0$ and $\tilde{q}_1$ are the parameters for spectral recovery. In [16], 1.18 and -0.09 were obtained as the values of $\tilde{q}_0$ and $\tilde{q}_1$.

2.3 PLATINUM: Aperiodic parameter extraction algorithm [17]

Mixed excitation [27] and aperiodicity [28] have usually been used to synthesize natural speech. In Legacy-STRAIGHT and TANDEM-STRAIGHT, aperiodicity is used as a speech parameter for synthesizing both periodic and aperiodic signals [29], [30]. WORLD takes another approach, one based on PLATINUM [17]. Legacy-STRAIGHT and TANDEM-STRAIGHT use the aperiodicity, whereas WORLD uses the excitation signal directly calculated from the waveform, F0, and spectral envelope.

PLATINUM windows the waveform by using a window with a length of $2T_0$. The spectrum of the windowed signal $X(\omega)$ is divided by the minimum phase spectrum $S_m(\omega)$. $S_m(\omega)$ is calculated as follows.

$$S_m(\omega) = \exp\left(\mathcal{F}[c_m(\tau)]\right), \quad (7)$$

$$c_m(\tau) = \begin{cases} 2c(\tau) & (\tau > 0) \\ c(\tau) & (\tau = 0) \\ 0 & (\tau < 0) \end{cases}, \quad (8)$$

$$c(\tau) = \mathcal{F}^{-1}[\log(P_l(\omega))], \quad (9)$$

The extracted excitation signal $x_p(t)$ is expressed as

$$x_p(t) = \mathcal{F}^{-1}[X_p(\omega)], \quad (10)$$

$$X_p(\omega) = \frac{X(\omega)}{S_m(\omega)}. \quad (11)$$

In PLATINUM, the temporal positions associated with each vocal cord vibration must be determined, and the F0 contour and waveform are used to determine them. First, the voiced section is determined; then, the temporal center position t_a of the section is determined. After that, an interval, $t_a \pm T_0$, is calculated, and the temporal position within it that has the maximum value of $y(t)^2$ is determined as the origin. Once the origin is determined, the vocoder-based synthesis algorithm automatically calculates the other vocal cord positions, defined as the origins of each impulse response on the basis of the F0 contour. This process is carried out on all voiced sections.

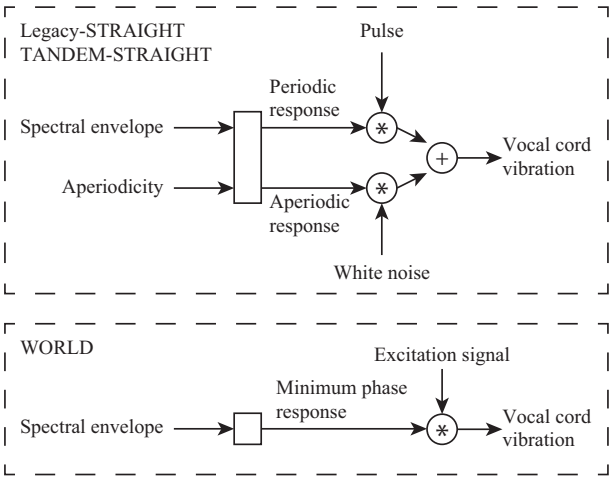


Fig. 3 Architecture for synthesizing a vocal cord vibration. The symbol $*$ represents convolution. Legacy-STRAIGHT uses group delay manipulation for periodic response synthesis.

2.4 Synthesis algorithm

Legacy-STRAIGHT and TANDEM-STRAIGHT calculate each vocal cord vibration independently from the periodic and aperiodic responses. TANDEM-STRAIGHT uses the periodic response directly, while Legacy-STRAIGHT manipulates the group delay to avoid buzzy timbre [31].

In WORLD, the vocal cord vibration is calculated on the basis of the convolution of the minimum phase response and the extracted excitation signal. Figure 3 illustrates the architecture of all the systems. WORLD has fewer convolutions than Legacy-STRAIGHT or TANDEM-STRAIGHT do, so its computational cost is lower. The F0 information is used to determine the temporal positions of the origin of each vocal cord vibration. Figure 4 illustrates examples of an input waveform and the excitation signals and synthesized waveforms of each system. In the cases of Legacy- and TANDEM-STRAIGHT, since the excitation signal depends on not only the spectral envelope but also the aperiodicity, it was calculated from the flattened spectral envelope. The synthesized waveforms of these systems are different, but the waveform synthesized by WORLD seems to be the most similar one to the input waveform.

2.5 Implementation

The system was implemented in C-language and in Matlab (both versions are available at our Website[†]). As the latest version of Legacy-STRAIGHT is only in Matlab, the evaluations described below used the Matlab versions of all the

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

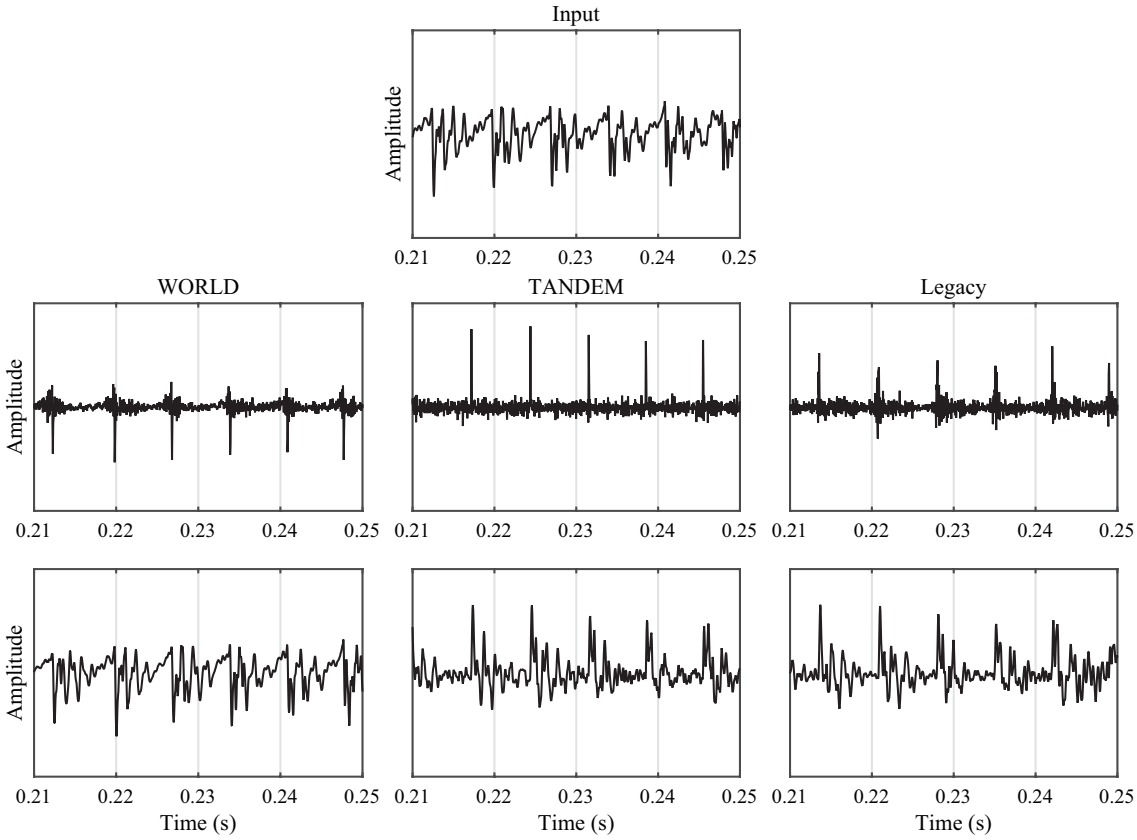


Fig. 4 Examples of input waveform, excitation signals, and synthesized waveforms. Top: Input waveform. Middle: Excitation signals before convolution. Bottom: Synthesized waveforms.

systems in order to make a fair comparison of their processing speeds.

3. Evaluation

The subjective evaluation was based on MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) [32], and it and the processing speed evaluation compared WORLD against Legacy-STRAIGHT and TANDEM-STRAIGHT. Legacy-STRAIGHT used nearly defect-free (NDF) F0 trajectory extraction [33], its most accurate F0 estimation algorithm.

3.1 Analysis and synthesis commands of each system

All of the systems have tuning parameters that the user can set in order to reduce the estimation error and improve performance. A brief optimization was carried out, and the results showed that tuning can make all of the systems perform better in specific situations. Since the tunings affect the quality of the synthesized speech, their influence should be avoided when making evaluations of the actual performance of the systems. Thus, no manual parameter tunings were carried out.

The version of WORLD was v0.1.4_2. The analysis and synthesis commands of this version are as follows.

1. *f0 = Dio(x, fs);*
2. *spec = CheapTrick(x, fs, f0);*
3. *source = Platinum(x, f0, spec);*
4. *y = SynthesisByWORLD(source, spec);*

where *x* represents the input waveform, and *fs* represents the sampling frequency. *y* represents the synthesized speech. In WORLD, since *source* includes both F0 and the excitation signal, the synthesis function only uses *source* and *spec*.

The version of TANDEM-STRAIGHT was TandemSTRAIGHTmonolithicPackage004TestRev. Its analysis and synthesis commands are as follows.

1. *f0 = exF0candidatesTSTRAIGHTGB(x, fs);*
2. *f0 = autoF0Tracking(f0, x);*
3. *f0.vuv = refineVoicingDecision(x, f0);*
4. *source = aperiodicityRatioSigmoid(x, f0, 1, 2, 0);*
5. *spec = exSpectrumTSTRAIGHTGB(x, fs, source);*
6. *y = exTandemSTRAIGHTsynthNx(source, spec);*

In TANDEM-STRAIGHT, the result of first estimation is modified by the other algorithm. *vuv* represents the voiced/unvoiced information. *source* includes both F0 and the aperiodicity. The synthesis function therefore uses *source* and *spec*. The arguments used for the *aperiodici-*

Table 1 Analysis conditions for each method. Lower and upper limits are the parameters for the F0 estimation and define the search range.

	Method		
	WORLD	TANDEM	Legacy
Frame shift (ms)	5	5	1
Lower limit (Hz)	80	32	40
Upper limit (Hz)	640	650	800
FFT size (sample)	2,048	2,048	2,048

Table 2 Characteristics of the speech used in the evaluation.

Number of speakers	4 (2 males and 2 females)
Number of speech	40 (10 words per speaker)
Kind of speech	4-mora word
Sampling / Quantization	48 kHz / 16 bit
Total length	32.18 s

tyRatioSigmoid function were the default parameters. The series of commands and the parameters were the same as those in the tutorial of TANDEM-STRAIGHT.

The version of Legacy-STRAIGHT was STRAIGHTV40_006b. The analysis and synthesis commands are as follows.

1. *[f0, ap] = exstraightsource(x, fs);*
2. *spec = exstraightspec(x, f0, fs);*
3. *y = exstraightsynth(f0, spec, ap, fs);*

In Legacy-STRAIGHT, F0 and aperiodicity are estimated at the same time.

Table 1 lists the common parameters for speech analysis. The frame shift of Legacy-STRAIGHT is different from those of TANDEM-STRAIGHT and WORLD. The lower and upper limits are the parameters for the F0 estimation and are used to define the search range. There are many parameters in each algorithm; the default parameters were used in each case in order to evaluate only the actual performance of the systems.

3.2 Characteristics of speech used in the evaluation

A database consisting of four-mora words was used for the evaluation. The conditions are listed in Table 2. In particular, the speech included consonants. This is in contrast to past evaluations of conventional systems that used speech consisting of only vowels [16], [17].

Figure 5 illustrates the cumulative distribution calculated from all F0 contours. The F0 contour used for the graph was estimated by NDF, which could cover widest range of F0. This graph shows that almost all the F0s were from 100 to 300 Hz.

3.3 Subjective evaluation of sound quality

Table 3 lists the conditions of the subjective evaluation. Ten subjects with normal hearing ability participated. Each subject used a GUI to rate the score. A snapshot of the GUI is shown in Fig. 6. Each evaluation set consisted of a piece of original speech and four pieces of anonymous speech, which consisted of a piece of original speech and three pieces of

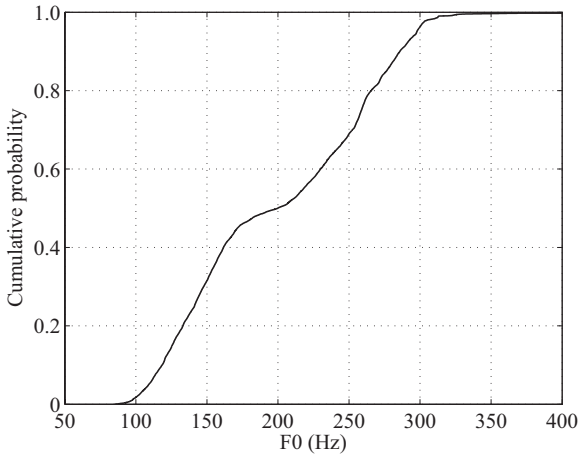


Fig. 5 Cumulative distribution of F0 used in the evaluation.

Table 3 Conditions of the subjective evaluation.

Method	MUSHRA
Number of subjects	10 students
Environment	28.9 dB (A-weighted SPL)
Headphones	SENNHEISER HD650
Audio I/O	Roland QUAD-CAPTURE

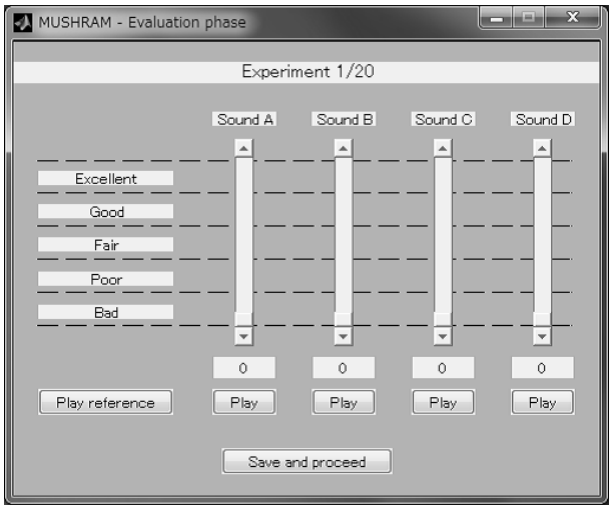


Fig. 6 Snapshot of the GUI used in the evaluation.

speech synthesized with WORLD, TANDEM-STRAIGHT, and Legacy-STRAIGHT. Since the evaluation was based on MUSHRA, subjects were instructed to give full marks (100 points) to at least one piece of speech.

Figure 7 illustrates the results. The vertical axis represents the MUSHRA score associated with the sound quality. The error bar represents the 95% confidence interval. The results of all speech show that the developed system was superior to the others. There were significant differences between all combinations (all p-values are under 0.0001).

The synthesized male speech showed the largest significant differences. Specifically, male speech synthesized with

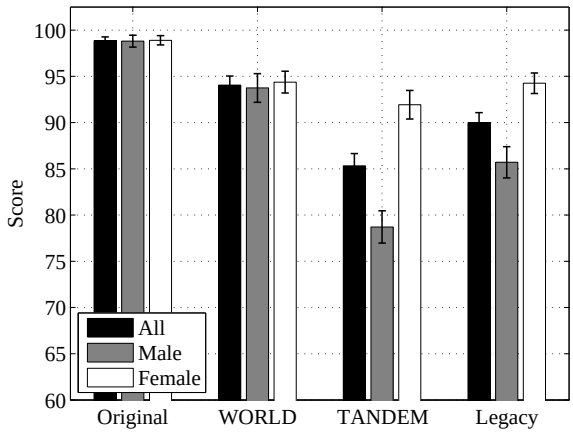


Fig. 7 Results of the subjective evaluation.

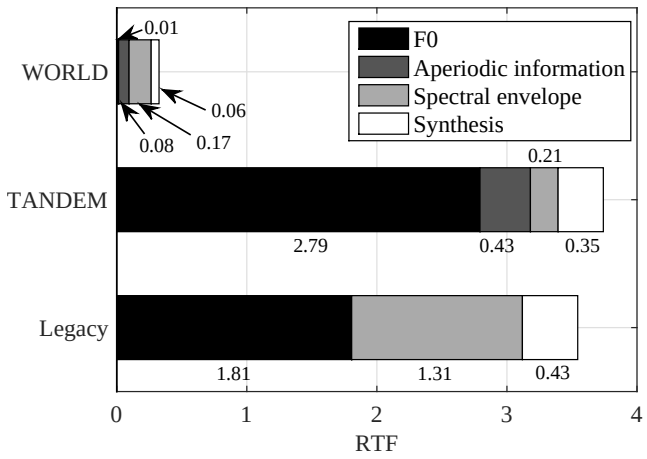


Fig. 8 Results of RTF. Note that the results for the Legacy-STRAIGHT implementation merge the F0 and aperiodicity estimations.

TANDEM-STRAIGHT and Legacy-STRAIGHT had lower sound quality than that of female speech. The differences in the female speech were therefore relatively smaller. In particular, there was no significant difference between WORLD and Legacy-STRAIGHT for female speech. These results suggest that WORLD was superior to the other systems on male speech and as good as Legacy-STRAIGHT on female speech.

3.4 Evaluation of processing speed

The systems were evaluated in terms of the real time factor (RTF). A mobile PC (Intel Core i7-3540M CPU 3.00 GHz, and 16.0 GB RAM) was used, and the version of Matlab was R2013a. All systems were implemented in Matlab, and no parallel processing was used in the evaluation. Note that the analysis parameters varied among the systems. In particular, frame shift directly affects the processing speed, and the smaller value was used in Legacy-STRAIGHT.

Figure 8 illustrates the results. The horizontal axis represents the RTF. The RTF is 1 if an input signal lasting n s was processed in n s. Legacy-STRAIGHT estimated the F0 and aperiodicity at the same time, so its F0 plot contains both results. The results in this case clearly show that only WORLD had an RTF that reflected a real-time processing capability.

The algorithms of WORLD were superior to the algorithms of the other systems in terms of processing speed. In Legacy-STRAIGHT, since the frame shift was set to 1 ms, the RTF of the spectral envelope estimation was much worse than those of the other algorithms. However, even if the frame shift was set to 5 ms, its RTF was still the worst.

4. Discussion

The results indicated that WORLD is capable of not only high-quality speech synthesis but also real-time processing. In this section, we discuss the system's effectiveness.

4.1 Sound quality of synthesized speech

The speech synthesized with WORLD had the best sound quality of all the synthesized speech. For example, the frame shift of Legacy-STRAIGHT was set to 1 ms, whereas that of WORLD was set to 5 ms. Furthermore, Legacy-STRAIGHT also used the group delay manipulation in synthesis as the technique for improving sound quality. Despite such an adverse condition, WORLD performed the best.

On the other hand, since DIO requires high-SNR speech, WORLD cannot synthesize natural speech from speech including additive noise. In particular, voiced/unvoiced estimation errors may degrade its sound quality. Thus, an important task for the future is to improve the noise robustness of WORLD.

The qualitative evaluation was carried out on the subjective listening tests. The differences in sound quality between the original speech and that of WORLD appeared at the boundaries of the phonemes. Moreover, the speech synthesized by WORLD differed from that of the other systems not only at the phoneme boundaries but also in terms of the sound quality of vowels. Specifically, the vowels of male speech synthesized with TANDEM-STRAIGHT were rated worse than those of the other systems. The main reason would be the phase difference in synthesized speech. Humans can easily perceive phase differences in speech with lower F0 compared with speech with higher F0 [34], and this suggests that an approximation using the minimum phase is inappropriate for low-pitch speech. Since the group delay manipulation of Legacy-STRAIGHT can improve the sound quality of low-pitch speech, this manipulation may have been one of the factors contributing to the difference between the results of TANDEM-STRAIGHT and Legacy-STRAIGHT. Phase information is currently being used [35], and the results suggest that it can improve sound quality. In the future, we should try to improve WORLD by incorpo-

rating an efficient phase modeling.

The robustness of the F0 and spectral envelope manipulation has already been evaluated [17]. However, it is difficult to discuss the quality of pitch-shifted speech for unvoiced speech. In particular, since the systems of the evaluation used different aperiodic parameters, the sound quality depended on the algorithm for manipulating F0. Pitch manipulation algorithms for each system need to be developed, and another subjective evaluation with pitch-shifted speech should be carried out.

Several performance indices have been proposed for evaluating spectral envelopes. However, the most popular ones, i.e., perceptual evaluation of speech quality (PESQ) and log-spectral distance (LSD), are useful for telephony system evaluations, but not appropriate for high-quality speech synthesizers. This situation points to the need for a new evaluation index for speech sampled at a high sampling rate.

4.2 Processing speed

The experiments demonstrated that WORLD was superior to other systems in terms of processing speed. Here, let us discuss its advantages in this regard. The F0 estimations of the other systems require a computationally intensive short time Fourier transform (STFT) in each frame. On the other hand, DIO filters the whole waveform and calculates the zero-crossing intervals much faster than STFT. Moreover, the other systems use two power spectra to calculate the spectral envelope, while CheapTrick uses only one. In the aperiodic parameter estimation, TANDEM-STRAIGHT uses an inverse matrix, while Legacy-STRAIGHT requires post-processing after calculating with a 1-ms frame shift. Although it is difficult to compare the computational costs of WORLD and TANDEM-STRAIGHT, PLATINUM has the advantage that it requires no post-processing, unlike Legacy-STRAIGHT. Finally, WORLD can synthesize speech by making a simple convolution, as shown in Fig. 3.

All of the systems were implemented in Matlab, and their processing speeds would improve if they were implemented in other computer languages such as C. Since the F0 estimations of WORLD and TANDEM-STRAIGHT can be parallelizable, using a graphics processing unit (GPU) would dramatically increase the processing speed. WORLD and TANDEM-STRAIGHT already have been implemented in C, and their processing speeds need to be optimized. Increasing the processing speed of the analysis and synthesis parts would enable us to develop a real-time voice conversion system. Such systems require processing for the parameter conversion, and a reduction in computational costs would give leeway for this conversion.

The processing speeds of the synthesis parts of all systems depended on F0 because their algorithms synthesized each vocal cord vibration. The use of a real-time synthesizer such as Vocaine [36] would be another way to improve WORLD.

5. Conclusions and future work

A high-quality speech synthesis system, named WORLD, was developed. WORLD consists of modern algorithms for estimating speech parameters. In a series of evaluations, the quality of sound synthesized with it was found to be superior to that of other synthesis systems. WORLD also worked ten times faster than the other systems.

We plan to use WORLD for voice conversion purposes such as voice morphing [37] and statistic parametric speech synthesis [38]. Its real-time applications include a singing synthesizer. WORLD is now available in the form of C and Matlab implementations. To improve convenience, we will implement it on other platforms. We will also develop Web APIs and embedded systems using digital signal processing (DSP) and field-programmable gate arrays (FPGAs).

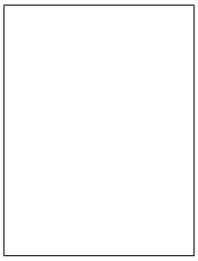
Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 15H02726 and 26540087 and the Research Institute of Electrical Communication, Tohoku University (H25/A08).

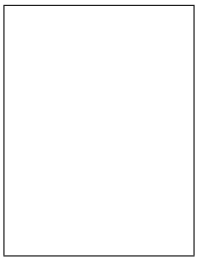
References

[1] H. Kenmochi, “Singing synthesis as a new musical instrument,” in Proc. ICASSP2012, pp.5385–5388, 2015.
[2] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Improvements of the one-to-many eigenvoice conversion system,” IEICE Trans. on Information and Systems, vol.E93-D, no.9, pp.2491–2499, 2010.
[3] H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol.11, no.2, pp.169–177, 1939.
[4] A.W. Black and N. Campbell, “Optimising selection of units from speech databases for concatenative synthesis,” in Proc. EUROSPEECH95, vol.1, pp.581–584, 1995.
[5] H. Kawahara, I. Masuda-Katsuse, and A. 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.
[6] J.L. Flanagan and R.M. Golden, “Phase vocoder,” Bell System Technical Journal, vol.45, no.9, pp.1493–1509, 1966.
[7] 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.
[8] R. McAulay and T.F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” IEEE Trans. on Acoustics, Speech and Signal Processing, vol.34, no.4, pp.744–754, 1986.
[9] K. Nakano, M. Morise, and T. Nishiura, “Vocal manipulation based on pitch transcription and its application to interactive entertainment for karaoke,” Lecture Notes in Computer Science, vol.LNCS 6851, pp.52–60, 2011.
[10] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, “Implementation of realtime straight speech manipulation system,” Acoust. Sci. & Tech., vol.28, no.3, pp.140–146, 2007.
[11] 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.
[12] 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

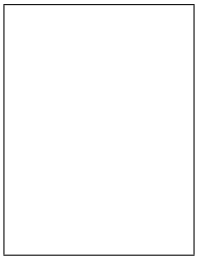
2008, pp.3933–3936, 2008.
[13] H. Kawahara and M. Morise, "Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework," SADHANA — Academy Proceedings in Engineering Sciences, vol.36, no.5, pp.713–728, 2011.
[14] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in Proc. AES 35th International Conference, CD-ROM Proceedings, 2009.
[15] M. Morise, H. Kawahara, and T. Nishiura, "Rapid f0 estimation for high-snr speech based on fundamental component extraction," IEICE Trans. on Information Systems, vol.J93-D, no.2, pp.109–117, 2010 (in Japanese).
[16] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," Speech Communication, vol.67, pp.1–7, 2015.
[17] M. Morise, "Platinum: A method to extract excitation signals for voice synthesis system," Acoust. Sci. & Tech., vol.33, no.2, pp.123–125, 2012.
[18] M. Morise, "Error evaluation of an f0-adaptive spectral envelope estimator in robustness against the additive noise and f0 error," IEICE trans. on Information Systems, vol.E98-D, no.7, pp.1405–1408, 2015.
[19] W. Hess, Pitch determination of speech signals, Springer-Verlag, 1983.
[20] A.M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal pitch detection," J. Acoust. Soc. Am., vol.36, no.2, pp.269–302, 1964.
[21] A. Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., vol.111, no.4, pp.1917–1930, 2002.
[22] A. Camacho and J.G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," J. Acoust. Soc. Am., vol.124, no.3, pp.1638–1652, 2008.
[23] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol.50, no.2B, pp.637–655, 1971.
[24] 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.
[25] M. Morise, T. Matsubara, K. Nakano, and T. Nishiura, "A rapid spectrum envelope estimation technique of vowel for high-quality speech synthesis," IEICE Trans. on Information Systems, vol.J94-D, no.7, pp.1079–1987, 2011. (in Japanese).
[26] M.V. Mathews, J.E. Miller, and E.E. David, "Pitch synchronous analysis of voiced sounds," J. Acoust. Soc. Am., vol.33, no.2, pp.179–185, 1961.
[27] A.V. McCree and T.P.B. III, "A mixed excitation lpc vocoder model for low bit rate speech coding," IEEE Transactions on Speech Audio Processing, vol.3, no.4, pp.242–250, 1995.
[28] D.W. Griffin and J.S. Lim, "A new model-based speech analysis/synthesis," in Proc. ICASSP1985, vol.10, pp.513–516, 1985.
[29] H. Kawahara and M. Morise, "Simplified aperiodicity representation for high-quality speech manipulation systems," in Proc. ICSP2012, pp.579–584, 2012.
[30] H. Kawahara, M. Morise, T. Toda, 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. Interspeech2014, pp.2243–2247, 2014.
[31] H. Kawahara, J. Estill, and O. Fujimura, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in Proc. ICASSP1997, pp.1303–1306, 1997.
[32] I.R.R. BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2003.
[33] H. Kawahara, A. Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight," in Proc. Interspeech2005, pp.537–540, 2005.
[34] R. Plomp and H.J. Steeneken, "Effect of phase on the timbre of complex tones," J. Acoust. Soc. Am., vol.46, no.2, pp.409–421, 1969.
[35] R. Maia, M. Akamine, and M. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in Proc. ICASSP2012, pp.4581–4584, 2012.
[36] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in Proc. ICASSP2015, pp.4230–4234, 2015.
[37] H. Kawahara, M. Morise, H. Banno, and V.G. Skuk, "Temporally variable multi-aspect n-way morphing based on interference-free speech representations," in Proc. APSIPA ASC 2013, pp.1–10, 2013.
[38] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," Speech Communication, vol.51, no.11, pp.1039–1064, 2009.



Masanori Morise received the Ph.D. degree in engineering from Wakayama University in 2008. He was a JSPS Research Fellow (DC1) in 2006–2008, a postdoctoral researcher at Kwansei Gakuin University in 2008–2009, and an Assistant Professor at Ritsumeikan University in 2009–2013. He is currently a Project Assistant Professor at University of Yamanashi. His research interests include speech analysis/synthesis and speech perception.



Fumiya Yokomori received the B.E. degree in Computer and Media Engineering from the University of Yamanashi in 2015. He is currently belonging to the Graduate School of Medicine and Engineering Science Department of Education. His research interests include speech analysis and psychological speech perception.



Kenji Ozawa received the B.E., M.E., and Ph.D. degrees in Electrical Communication from Tohoku University in 1986, 1988, and 1994, respectively. He is currently a professor of the University of Yamanashi. His research interests include psychoacoustics, signal processing of audio signals, and *Kansei* (Emotion and Sensibility) information processing.