ACOUSTICAL LETTER

Sound quality comparison among high-quality vocoders by using re-synthesized speech

Masanori Morise* and Yusuke Watanabe

Faculty of Engineering, University of Yamanashi,
4–3–11 Takeda, Kofu, 400–8511 Japan

(*Received 26 October 2017, Accepted for publication 15 January 2018*)

Keywords: Speech analysis, Speech synthesis, Vocoder, Subjective evaluation, Sound quality
PACS number: 43.72.Ar, 43.72.Ja [doi:10.1250/ast.39.263]

1. Introduction

Speech analysis/synthesis systems have been used in various kinds of applications such as voice conversion [1] and statistical parametric speech synthesis [2]. These applications use a high-quality system based on a vocoder [3], and STRAIGHT [4] is one of the best ones. In this paper, "vocoder" means the speech analysis/synthesis system, and the high-quality vocoder accurately decomposes the speech waveform into the fundamental frequency (f_o), spectral envelope, and aperiodicity. In recent years, we have proposed a new vocoder named WORLD [5]. Both STRAIGHT and WORLD have been used in several applications such as the Merlin toolkit [6], and recently WORLD has been used in other applications [7,8].

Since we have released WORLD on GitHub* and have been continuously updating WORLD to improve the sound quality of the synthesized speech, there is no information on the performance of the current version of WORLD. The purpose of this study is to compare high-quality vocoders including STRAIGHT and the old and current versions of WORLD. To evaluate them, there are several approaches such as checking the sound quality after voice conversion and statistical parametric speech synthesis. In this paper, an evaluation by using re-synthesized speech was carried out to discuss the most basic performance. The difference among them and the characteristics of each vocoder are discussed by using the obtained result.

2. Subjective evaluation

In this section, we explain the protocol of the subjective evaluation.

2.1. Vocoders used for evaluation

Four vocoders were selected for comparison. In WORLD, the old and current versions were used. In f_o estimation, the old version used DIO [9], which is a fast and reliable f_o estimator. On the other hand, the current version used Harvest [10]. They used CheapTrick [11,12] and D4C [13] as the spectral envelope and aperiodicity estimators, respectively. We did not use WORLD (PLATINUM edition [5,14]) because it has a different representation in the aperiodic parameter.

STRAIGHT [4] was used as the highest quality vocoder, and NDF [15], which is the latest f_o estimator of STRAIGHT, was used. YANG VOCODER† was also used as a modern vocoder. We did not use TANDEM-STRAIGHT [16] because our previous study has shown that its sound quality is significantly inferior to that of STRAIGHT and WORLD [13]. There are speech analysis algorithms for achieving high-quality speech synthesis such as Nakano *et al.*'s one [17], but only the vocoders that have three estimators were selected in this evaluation.

2.2. Difference between old and current versions of WORLD

The main difference between them is the estimation algorithm in f_o and the aperiodicity. Harvest attempts to reduce the unvoiced segment and gives the reliable f_o to the segment for continuous F0 modeling [18]. In cases where the unvoiced segment is wrongly identified as the voiced segment, the aperiodicity estimated by D4C often causes degradation of the sound quality. The aperiodicity of the unvoiced segment must be 1.0 in the whole frequency band because the whole component of the spectral envelope comes from the aperiodic component. D4C occasionally gives a low value in the lower frequency band, and as a result, the periodic component is perceived as the noise. The current version adds a process in D4C to identify the voiced/unvoiced segment and give the value of 1.0 in the whole frequency band in cases where a frame has an f_o but is identified as the unvoiced segment. This process is called *D4C LoveTrain* in the source code of WORLD.

Since the voiced segment contains a vocal cord vibration that has a -6 dB/oct slope in the power spectrum, the power ratio between the lower and higher frequency bands is effective to identify whether the segment contains a vocal cord vibration. Power from 100 to 4,000 Hz and power from 100 to 7,900 Hz are used as the lower and higher frequency bands, respectively. Power ratio c is given by the following equation.

$$c = \int_{100}^{4000} P(f)df \Bigg/ \int_{100}^{7900} P(f)df, \tag{1}$$

where $P(f)$ represents the power spectrum, and f represents the frequency (Hz). The current version of WORLD uses a c of 0.85 as the threshold, and the frame with a c below 0.85 is identified as the unvoiced segment. The aperiodicity of this frame is set to 1.0 in the whole frequency band. Since DIO

*e-mail: mmorise@yamanashi.ac.jp
https://github.com/mmorise/World

†https://github.com/google/yang_vocoder

Table 1 Speech used for evaluation.

Number of speakers	Four (two men and two women)
Sampling	48 kHz/16 bit
Number of speeches	40 (10 per speaker)
Kinds of speech	4-mora words including consonants

Table 2 Experimental conditions.

Method	MUSHRA
Number of subjects	Fourteen
Environment	Soundproof room
A-weighted SPL	18 dB
Audio I/O	Roland QUAD-CAPTURE
Headphones	SENNHEISER HD650

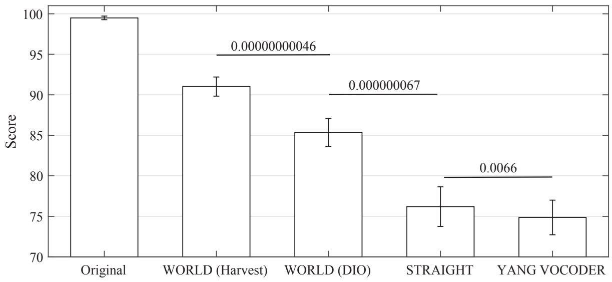

Fig. 1 Results of subjective evaluation. Value over horizontal line represents adjusted p-value.

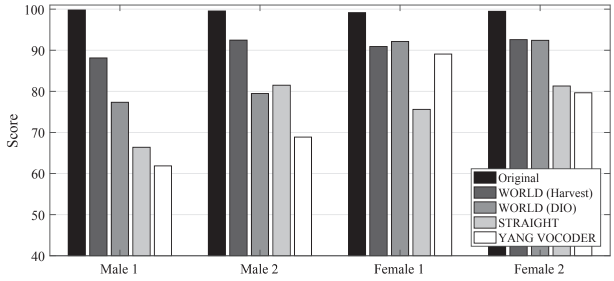

Fig. 2 Results of each speaker.

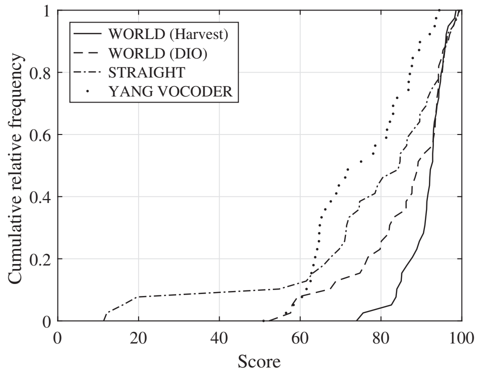

Fig. 3 Cumulative relative frequency distributions of each result. Horizontal and vertical axes represent score and cumulative possibility, respectively.

includes the accurate voiced/unvoiced detection, this process is skipped.

2.3. Evaluation method

A MUSHRA-based evaluation [19] was carried out to compare the sound quality of each vocoder. The speech used for the evaluation was shown in Table 1. 40 speech waveforms were randomly selected from speech database FW07 [20]. We did not use long sentences to accurately evaluate the degradation caused by an error in short period. In all the vocoders, the frame shift was set to 1 ms which is the default value of STRAIGHT, and their default values were used in other parameters such as the floor and ceiling frequencies for f_0 estimation.

Table 2 shows the experimental conditions in the evaluation. A soundproof room with an A-weighted SPL of 18 dB was used for the evaluation. Fourteen subjects with normal hearing ability attended the evaluation. The sound stimuli were reproduced through the headphones, and the sound pressure level was set to not exceed 70 dB.

3. Results and discussion

Figure 1 illustrates the evaluation results. The vertical axis represents the MUSHRA score corresponding to the sound quality. The error bars represent the 95% confidence interval. The value over each horizontal line represents the adjusted p-value.

In the statistical analysis, we used the Wilcoxon signed-rank test because not all populations could be assumed to be normally distributed. The adjusted p-values were calculated based on the Bonferroni correction. We skipped the comparison between original and re-synthesized speech because the difference of sound quality was obvious. Since only multiple comparisons among vocoders were carried out, the number of pairs was six. Therefore, the adjusted p-value was calculated to be six times the raw p-value. Several p-values with the obvious difference were omitted from the figure. For example, we omitted the result between WORLD (Harvest) and STRAIGHT because their difference is larger than that between WORLD (Harvest) and WORLD (DIO).

The results showed that the WORLD (Harvest) was significantly superior to the others in sound quality. WORLD (DIO) was the best vocoder compared with the STRAIGHT and YANG VOCODER. STRAIGHT was significantly superior to YANG VOCODER.

To discuss the characteristics of each vocoder, we analyzed the experimental results in each speaker. Figure 2 illustrates the evaluation results, which are separately calculated in each speaker. WORLD (Harvest) can synthesize natural speech from all speakers. Compared with the others, it was difficult for YANG VOCODER to synthesize natural speech from male speakers. Since a buzzy timbre in vowels was often observed, the main cause seems to be the accuracy of the spectral envelope.

To discuss the trend, the cumulative relative frequency distributions of each vocoder are shown in Fig. 3. This figure shows that the scores of four speech synthesized by STRAIGHT fell below 20. The main cause was the error that the voiced segment was wrongly identified as the unvoiced segment. YANG VOCODER could accurately estimate f_o from all speech waveforms, but the sound quality was relatively bad. The buzzy timbre of synthesized speech was the main reason. In comparison between STRAIGHT and YANG VOCODER, the significant difference was observed even if it seems that there was not enough difference. The cause was that the difference between average scores was not large, but that between median scores was enough to show the significant difference.

In comparison between WORLD (Harvest) and WORLD (DIO), the difference of sound quality was observed at the boundary of voiced/unvoiced segments. This difference suggests that the combination of Harvest and D4C LoveTrain can work as expected. In short, the result clearly showed that the current version of WORLD was the best of all vocoders.

4. Conclusion

This paper showed the difference among several high-quality vocoders. The MUSHRA-based evaluation result showed that the current version of WORLD could achieve the best performance. In the analysis of each speech, there are several speech waveforms for which STRAIGHT cannot estimate the f_o. YANG VOCODER could not totally achieve natural speech because of the low-accuracy of the spectral envelope. In comparison between the old and current versions of WORLD, the current version was superior to the old one.

The next goal is a comparison among them in the voice conversion and the statistical parametric speech synthesis. Since WORLD cannot synthesize speech that is as natural as the input, improvement of the sound quality is also important.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP15H02726, JP16H05899, JP16K12511, and JP16K12464.

References

[1] T. Toda, A. W. Black and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, **15**, 2222–2235 (2007).

[2] H. Zen, K. Tokuda and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, **51**, 1039–1064 (2009).

[3] H. Dudley, "Remaking Speech," *J. Acoust. Soc. Am.*, **11**, 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 Commun.*, **27**, 187–207 (1999).

[5] M. Morise, F. Yokomori and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, **E99-D**, 1877–1884 (2016).

[6] Z. Wu, O. Watts and S. King, "Merlin: An open source neural network speech synthesis system," *Proc. 9th ISCA Speech Synthesis Workshop* (*SSW9*), pp. 218–223 (2016).

[7] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," *Proc. Interspeech 2017*, pp. 4001–4005 (2017).

[8] Y. Taigman, L. Wolf, A. Polyak and E. Nachmani, "Voice synthesis for in-the-wild speakers via a phonological loop," *arXiv:1707.06588 [cs.LG]*, pp. 1–11 (2017).

[9] 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," *Proc. AES 35th International Conference*, 4 pages (2009).

[10] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," *Proc. Interspeech 2017*, pp. 2321–2325 (2017).

[11] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, **67**, 1–7 (2015).

[12] M. Morise, "Error evaluation of an F0-adaptive spectral envelope estimator in robustness against the additive noise and F0 error," *IEICE Trans. Inf. Syst.*, **E98-D**, 1405–1408 (2015).

[13] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, **84**, 57–65 (2016).

[14] M. Morise, "PLATINUM: A method to extract excitation signals for voice synthesis system," *Acoust. Sci. & Tech.*, **33**, 123–125 (2012).

[15] 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," *Proc. Interspeech 2005*, pp. 537–540 (2005).

[16] 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," *Proc. ICASSP 2008*, pp. 3933–3936 (2008).

[17] T. Nakano and M. Goto, "A spectral envelope estimation method based on F0-adaptive multi-frame integration analysis," *Proc. SAPA-SCALE 2012*, pp. 11–16 (2012).

[18] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio Speech Lang. Process.*, **19**, 1071–1079 (2011).

[19] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems" (2003).

[20] T. Kondo, S. Sakamoto, S. Amano and Y. Suzuki, "Compensation for list-difference of word intelligibility by conditioning signal-to-noise ratio: Validation by using the familiarity-controlled word lists 2007 (FW07)," *J. Acoust. Soc. Jpn.* (*J*), **69**, 224–231 (2013) (in Japanese).