INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

TUSK: A framework for overviewing the performance of F0 estimators

Masanori Morise[1], *Hideki Kawahara*[2]

[1]Interdisciplinary Graduate School, University of Yamanashi, Japan
[2]Faculty of Engineering, Wakayama University, Japan
mmorise@yamanashi.ac.jp, kawahara@sys.wakayama-u.ac.jp

Abstract

This article presents a framework for overviewing the performance of fundamental frequency (F0) estimators and evaluates its effectiveness. Over the past few decades, many F0 estimators and evaluation indices have been proposed and have been evaluated using various speech databases. In speech analysis/synthesis research, modern estimators are used as the algorithm to fulfill the demand for high-quality speech synthesis, but at the same time, they are competing with one another on minor issues. Specifically, while all of them meet the demands for high-quality speech synthesis, the result depends on the speech database used in the evaluation. Since there are various types of speech, it is inadvisable to discuss the effectiveness of each estimator on the basis of minor differences. It would be better to select the appropriate F0 estimator in accordance with the speech characteristics. The framework we propose, TUSK, does not rank the estimators but rather attempts to overview them. In TUSK, six parameters are introduced to observe the trends in the characteristics in each F0 estimator. The signal is artificially generated so that six parameters can be controllable independently. In this article, we introduce the concept of TUSK and determine its effectiveness using several modern F0 estimators.

Index Terms:Speech analysis, fundamental frequency, temporal variation, noise robustness

1. Introduction

Fundamental frequency (F0) is one of the most important parameters for speech processing. Speech synthesizers come with a standard function for manipulating F0. Although many F0 estimators have been proposed over the years, as yet there is no perfectly ideal algorithm because of the many conditions in speech. Speech often contains noise depending on the recording environment; for example, vibrato singing has an F0 contour that includes temporal fluctuation. The F0 of a periodic signal is defined as the smallest period of the vocal cord vibrations, and F0 estimators assume that speech is periodic in the short term. However, vibrato singing does not fulfill this assumption even if the period of analysis is short, so it is difficult to estimate an accurate F0 contour from such singing. Since real speech can become degraded for a variety of reasons, a perfect F0 estimator is the ultimate target.

In speech analysis/synthesis, speech used as the input is usually recorded in a silent environment, which is in contrast to the speech used for other studies such as those involving automatic speech recognition. A vocoder-based synthesizer [1] requires not only the F0 but also the spectral envelope, and F0 information is useful to accurately estimate this [2, 3, 4]. Speech analysis/synthesis research therefore tends to give higher priority to estimation accuracy than to noise robustness. On the other hand, noise robustness is important in speech processing using speech recorded in a real environment including noise. The appropriate F0 estimator therefore depends on the purpose of the study. It is important to not only rank the F0 estimator but also to overview the characteristics of F0 estimators.

In light of the above, we introduce a framework for overviewing the characteristics of F0 estimators. This framework, named TUSK, utilizes an artificial signal for the evaluation. The equation used for designing the signal has six parameters for measuring the characteristics of an F0 estimator. In this article, we explain the concept of TUSK and the six parameters. A computational simulation with several modern F0 estimators is carried out to demonstrate the effectiveness of TUSK.

The rest of this paper is organized as follows. In Section 2, we briefly discuss the conventional research on F0 estimators and methods for their evaluation. In Section 3, we propose our framework, TUSK. In Section 4, we evaluate the proposed framework with several modern F0 estimators. We conclude in Section 5 with a brief summary and a mention of future work.

2. Conventional F0 estimators and evaluation methods

Many F0 estimators based on acoustic features have been proposed [5]. In terms of acoustic features in the time domain, autocorrelation [6] is standard, and several improved algorithms such as YIN [7] and pYIN [8] have been proposed. Since the power spectrum of input speech has a harmonic structure, an acoustic feature based on the power spectrum is used for estimation. Cepstrum [9, 10] is standard, and SWIPE [11] has recently been proposed as an accurate estimator. In terms of other acoustic features, the instantaneous frequency [12] and events caused by vocal fold vibrations [13] are used as effective acoustic features. We have also proposed an algorithm based on fundamental component extraction in the harmonic structure [14]. Several algorithms specializing in noise robustness have also been proposed [15, 16]. Since the appropriate F0 estimator can be selected in accordance with the specific speech characteristics under consideration, it is important to understand the characteristics of each F0 estimator.

Performance evaluation has been carried out on real speech including an electroglottography (EGG) signal. The CMU ARCTIC database[1] and Paul Bagshaw's database[2] are usually used in such evaluation. The target F0 contour is calculated from the EGG signal and the estimation performance is then calculated from the difference between the target and estimated F0 contours. Various evaluation indices have also been proposed, such as fine pitch error (FPA), gross pitch error (GPA)

[1] http://festvox.org/cmu_arctic/index.html
[2] http://www.cstr.ed.ac.uk/research/projects/fda/

http://dx.doi.org/10.21437/Interspeech.2016-140

[17], and gross error [7].

One of the major problems is the reliability of the target F0 contour. An estimator is required to calculate the target F0 contour because the EGG signal is equivalent to the information in vocal cord vibrations. The F0 estimator used for estimating the target F0 contour affects the evaluation result—sometimes positively, sometimes negatively. To address this issue, TUSK uses an artificial signal generated from the target F0 contour. Conventional error indices are useful but they make it more difficult to simply discuss the performance, so TUSK uses the root mean square (RMS) error between the target and estimated F0 contours.

3. TUSK: The proposed framework

TUSK measures the relationship between estimation performance and six parameters. These six parameters are used to design the artificial signal used in the evaluation.

3.1. Signal commonly used in the evaluation

We first explain how to design the artificial signal that has six parameters. Signal $x(t)$ is a complex tone and given by

$$x(t) = n(t) + h(t) * \sum_{k=1}^{K} a_k \cos\left(2\pi k \int_0^t f_0(\tau)d\tau + \theta_k\right), \tag{1}$$

where $n(t)$ represents additive noise, $h(t)$ represents an impulse response, $*$ represents the convolution, and $f_0(t)$ represents an F0 contour. a_k and θ_k represent the amplitude and phase at k-th harmonic components, respectively. K represents the number of harmonics and is determined such that $Kf_0(t)$ does not exceed the Nyquist frequency. The influence of a parameter on the estimation performance is evaluated by controlling the parameter.

The average $f_0(t)$ is fixed to a basic F0 f_c. Note that since the F0 contour of real speech contains small fluctuation, a fluctuation model proposed by Klatt [18] is added to the basic F0.

$$\begin{aligned}\Delta f_0(t) &= \frac{FL}{50}\frac{f_c}{100}(\sin(2\pi 12.7t) + \sin(2\pi 7.1t) \\ &\quad + \sin(2\pi 4.7t)),\end{aligned} \tag{2}$$

where FL is the parameter associated with the flatter and is fixed to 25 in accordance with reference [18]. The target F0 contour $f_0(t)$ is defined as $f_0(t) = f_c + \Delta f_0(t)$. The basic parameters are defined as $n(t) = 0$, $h(t) = \delta(t)$, $a_k = 1$, $\theta_k = 0$, and $f_c = 440$.

3.2. ACT 1: Relationship between basic F0 and estimation performance

The first evaluation is carried out to confirm the frequency range in which the estimator can estimate an accurate F0. The parameter used in ACT 1 is the basic F0 f_c, and we can measure the frequency range by controlling it. Since the signal does not contain additive noise or reverberation, the result in ACT 1 generally achieves the highest performance.

3.3. ACT 2: Influence of temporal fluctuation in F0 contour

ACT 2 uses an F0 contour that has a parameter for measuring the influence of the temporal fluctuations. The additive F0 contour $f_v(t)$ is given by

$$f_v(t) = \sqrt{\alpha f_c}\cos\left(\sqrt{\alpha f_c}t\right), \tag{3}$$

where α represents the intensity of temporal fluctuation in the F0 contour and the maximum tilt indicates αf_c. This evaluation uses the F0 contour defined as $f_0(t) = f_c + \Delta f_0(t) + f_v(t)$. In this paper, we call the parameter α *vibrato intensity*. ACT 2 enables us to observe the influence of vibrato intensity on the estimation performance by controlling α.

3.4. ACT 3: Influence of amplitudes of each harmonic component

Algorithms focusing on the harmonic structure of a power spectrum would be weak against the variation of amplitude of each harmonic component. To determine the influence, a_k is randomized in ACT 3, and the dynamic range of randomization is used as the parameter. This dynamic range is given as the logarithmic amplitude, and uniform random numbers are used for randomization. This evaluation is repeated several times and its median value is used as the estimation performance.

3.5. ACT 4: Influence of phases of each harmonic component

In ACT 4, θ_k is randomized as with ACT 3. The phase difference between θ_k and θ_{k+1} affects the power by interference between neighboring harmonics. ACT 4 can measure the influence of the phase difference between neighboring harmonics. The parameter is the dynamic range, and its maximum value is 2π. Other conditions are the same as ACT 3.

3.6. ACT 5: Noise robustness

ACT 5 uses arbitrary noise, with the SNR as the parameter. Basic noise robustness is evaluated by the relationship between the SNR and the estimation performance. Since a_k is fixed to 1 in this evaluation, SNR in all frequency bands is fixed by using white noise. We can also use other types of noise to measure the influence of the noise type on the estimation performance.

3.7. ACT 6: Influence of the reverberation

The last evaluation measures the influence of the reverberation. In room acoustics, an impulse response is used to estimate the reverberation time T_{60}. The impulse response measured in a room contains early reflections and the reverberation. Since it is difficult to control them by one parameter, TUSK uses the impulse response designed by a simple parameter in the reverberation. The amplitude envelope is given by the following equation:

$$h_e(t) = \begin{cases} 0 & (t < 0) \\ 1 & (t = 0) \\ \dfrac{\exp\left(\dfrac{t\log(0.001)}{r}\right)}{\sqrt{10}} & (\text{otherwise}), \end{cases} \tag{4}$$

where r represents the parameter associated with the reverberation time T_{60}. The impulse response $h(t)$ is calculated by multiplying $h_e(t)$ by white noise $n(t)$. In the equation, the amplitude of the last term is set to $\sqrt{10}$ on the basis of early delay time (EDT), which is the reverberation time measured over the first 10 dB of the decay. This impulse response does not have early reflections; it only has the reverberation.

Table 1: Experimental conditions for each evaluation

Evaluation	Parameter
ACT 1	f_c: 2100...6900 cent
ACT 2	α: 0...25
ACT 3	a_k: 0...40 dB
ACT 4	θ_k: 0...2π
ACT 5 (white noise)	SNR: 0...60 dB
ACT 5 (pink noise)	SNR: 0...60 dB
ACT 6	r: 10...1000 ms

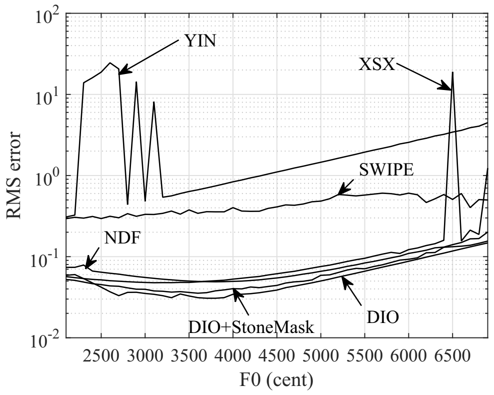

Figure 1: *Relationship between the basic F0 and the RMS error of each estimator. Results of YIN and XSX indicate locally high error.*

4. Evaluation

To determine the effectiveness of TUSK, we carried out a series of experiments by computer simulation.

4.1. F0 estimators used in the evaluation

We used several modern F0 estimators in the experiment. YIN [7] and SWIPE [11] were utilized as standard estimators focusing on the time and frequency domain, respectively. NDF [19] used in STRAIGHT [20] and XSX used in TANDEM-STRAIGHT [21, 22] are utilized as the high performance estimators. DIO [14] and StoneMask are also used for comparison. They are used in WORLD [23][3], which is a high-quality speech analysis/synthesis system. DIO requires high-SNR speech, and StoneMask is used to improve the noise robustness of the result estimated by DIO.

4.2. Common conditions

The length of the complex tone was set to 1.2 s and the sampling frequencies of $x(t)$ and $f_0(t)$ were 48 and 1 kHz, respectively. In all algorithms, the frame shift was set to 1 ms and the lower and upper limits in the F0 search were set to 40 and 1,000 Hz, respectively. Since several of the algorithms could not estimate the F0 of the head and tail, they were removed from the evaluation. The F0s of 1,000 samples ranging from 0.1 to 1.1 s were used to calculate the RMS error. The average RMS error was defined as the estimation performance of a condition.

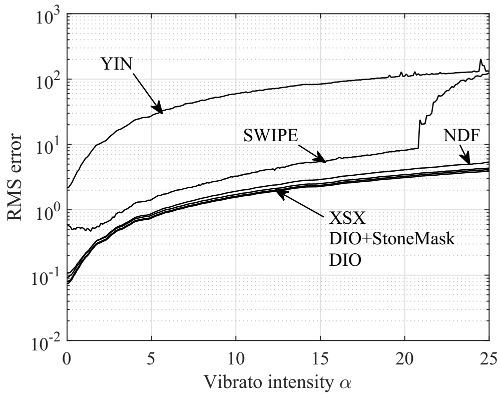

Figure 2: *Influence of vibrato intensity on the RMS error. The results of XSX, DIO, and DIO+StoneMask were almost all the same.*

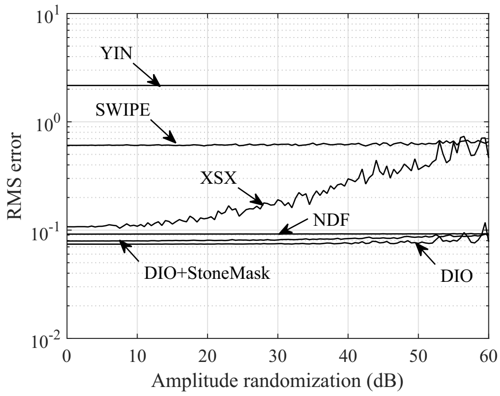

Figure 3: *Influence of the amplitude randomization on the RMS error.*

The conditions in the six parameters are shown in Table 1. ACT 5 used white and pink noises to confirm the difference of noise type. The number of iterations in ACTs 3, 4, 5, and 6 is 100, and the median value is used in the evaluation.

4.3. Results

Figure 1 shows the relationship between the basic F0 and the estimation performance. The horizontal and vertical axes represent the basic F0 and the RMS error, respectively. The RMS errors of YIN and SWIPE were higher than those of the others. The results of YIN and XSX have peaks in lower and higher F0, respectively. This can be attributed to typical F0 estimation errors such as half or double pitch errors.

Figure 2 shows the relationship between the vibrato intensity α and the RMS error. In all estimators, the estimation error increased in proportion to the vibrato intensity α. In SWIPE, the estimation error exponentially increased when the vibrato intensity exceeded 21. This suggests that SWIPE was inferior to the others in terms of temporal resolution.

Figures 3 and 4 show the influences of amplitude a_k and phase θ_k on the RMS error. In amplitude θ_k, XSX tends to increase the estimation errors in proportion to the dynamic range. The other estimators were not affected by the dynamic range. In phase θ_k, virtually non of the algorithms increased the estimation error. This suggests that the influence of phase difference on the power spectrum is small.

Figures 5 and 6 show the relationship between the SNR and the estimation performance. The results of white and pink noise, shown in Fig. 5 and 6, respectively, suggest that YIN and

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

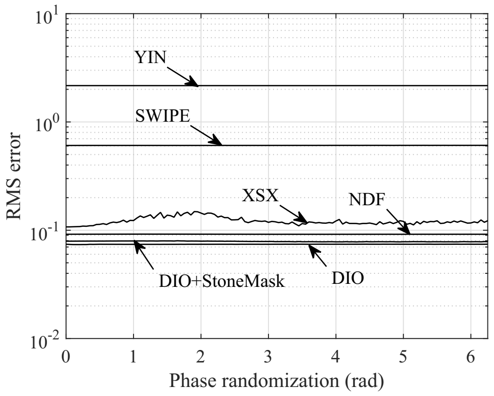

Figure 4: *Influence of the phase randomization on the RMS error.*

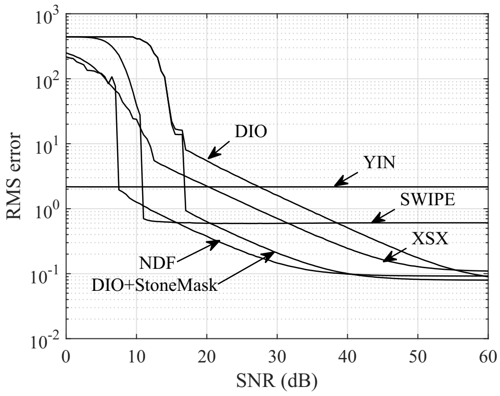

Figure 5: *Influence of the SNR on the RMS error (white noise).*

SWIPE were robust against the noise. DIO was the worst of all estimators in noise robustness, but DIO+StoneMask improved the robustness, as expected. Similar trends were observed between white and pink noise. In all estimators, the results suggest that the SNR in a lower frequency band is important.

Figure 7 shows the relationship between the reverberation parameter and the RMS error. YIN had a different trend than the other estimators. The results of the other estimators were almost all the same.

4.4. Discussion

The proposed framework was able to provide an overview of the characteristics of F0 estimators. For example, it showed that YIN was inferior to others in terms of estimation error but superior for noise robustness. Although NDF was superior in all parameters, it comes at a huge computational cost. In cases where fast processing speed is required, the DIO+StoneMask would be reasonable when the input speech is recorded in a silent environment. In speech recorded in a noisy environment, YIN achieves the best performance. Ultimately, the framework enables users to select the appropriate estimator on the basis of the result.

TUSK enabled us to discuss the characteristics of each F0 estimator. However, TUSK does not focus on processing speed or voiced/unvoiced estimation. Expansion of TUSK in the evaluation of voiced/unvoiced estimation and processing speed is important. To approximate the speech signal, another parameter such as amplitude modulation would be also important. Since current version of TUSK uses artificial signals that roughly approximate the real speech, an evaluation by speech database with EGG signal should also be carried out concurrently. A glottal source model based on the L-F model [24] can approximate real speech, and introduction of the model is one of the future works.

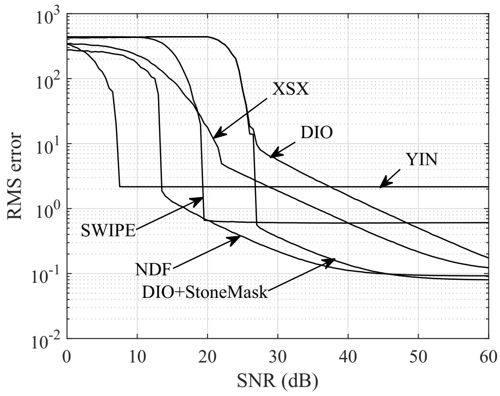

Figure 6: *Influence of the SNR on the RMS error (pink noise).*

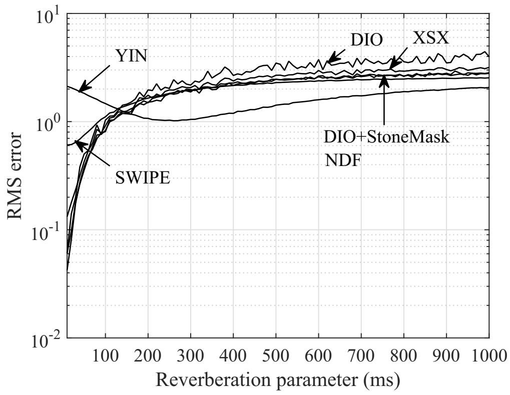

Figure 7: *Influence of the reverberation on RMS error. In this evaluation, the results of DIO+StoneMask and NDF were almost all the same.*

5. Conclusion

In this article, we introduced our framework for overviewing the performance of F0 estimators. Conventional research on F0 estimation has used several speech databases, but the proposed framework, TUSK, uses an artificial signal in which the target F0 contour is known. TUSK evaluates F0 estimators using six parameters.

The evaluation using several modern F0 estimators demonstrated that TUSK can evaluate the characteristics of each estimator, thus enabling users to select the appropriate F0 estimator for the characteristics of the target speech. Our next objective is to expand TUSK for the evaluation of other parameters. A system of for recommending an F0 estimator on the basis of target speech will also be key in our future work.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 26540087, 15H02726, 16H01734.

7. References

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

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

[3] ——, "Error evaluation of an f0-adaptive spectral envelope estimator in robustness against the additive noise and f0 error," *IEICE Trans. Inf. & Syst.*, vol. E98-D, no. 7, pp. 1405–1408, 2015.

[4] 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.

[5] W. Hess, *Pitch determination of speech signals*. Springer-Verlag, 1983.

[6] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Transactions on acoustic, speech, and signal processing*, vol. ASSP-22, no. 5, pp. 353–362, 1974.

[7] 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.

[8] M. Mauch and S. Dixon, "PYIN: a fundamental frequency estimator using probabilistic threshold distributions," in Proc. *ICASSP2014*, pp. 659–663, 2014.

[9] A. Noll, "Short-time spectrum and "cepstrum" techniques for vocal pitch detection," *J. Acoust. Soc. Am.*, vol. 36, no. 2, pp. 269–302, 1964.

[10] ——, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293–309, 1967.

[11] 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.

[12] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in Proc. *ICASSP95*, pp. 756–759, 1995.

[13] B. Yegnanarayana and K. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[14] M. Morise, H. Kawahara, and T. Nishiura, "Rapid f0 estimation for high-snr speech based on fundamental component extraction," *IEICE Trans. Inf. & Syst.*, vol. J93-D, no. 2, pp. 109–117, 2010 (in Japanese).

[15] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Transactions on speech and audio processing*, vol. 9, no. 7, pp. 727–730, 2001.

[16] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3690–3700, 2004.

[17] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on acoustic, speech, and signal processing*, vol. ASSP-24, no. 5, pp. 399–418, 1976.

[18] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 82, no. 2, pp. 820–857, 1990.

[19] 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.

[20] 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: Possible role of a reptitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[21] 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. *ICASSP2008*, pp. 3933–3936, 2008.

[22] 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.

[23] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. & Syst.*, vol. E99-D, no. 7, 2016.

[24] H. Kawahara, "SparkNG: interactive MATLAB tools for introduction to speech production, perception and processing fundamentals and application of the aliasing-free L-F model component," in Proc. *INTERSPEECH2016*, 2–page, 2016.