INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Harvest: A high-performance fundamental frequency estimator from speech signals

Masanori Morise[1]

[1]Faculty of Engineering, University of Yamanashi, Japan
mmorise@yamanashi.ac.jp

Abstract

A fundamental frequency (F0) estimator named *Harvest* is described. The unique points of Harvest are that it can obtain a reliable F0 contour and reduce the error that the voiced section is wrongly identified as the unvoiced section. It consists of two steps: estimation of F0 candidates and generation of a reliable F0 contour on the basis of these candidates. In the first step, the algorithm uses fundamental component extraction by many band-pass filters with different center frequencies and obtains the basic F0 candidates from filtered signals. After that, basic F0 candidates are refined and scored by using the instantaneous frequency, and then several F0 candidates in each frame are estimated. Since the frame-by-frame processing based on the fundamental component extraction is not robust against temporally local noise, a connection algorithm using neighboring F0s is used in the second step. The connection takes advantage of the fact that the F0 contour does not precipitously change in a short interval. We carried out an evaluation using two speech databases with electroglottograph (EGG) signals to compare Harvest with several state-of-the-art algorithms. Results showed that Harvest achieved the best performance of all algorithms.

Index Terms: speech analysis, fundamental frequency, fundamental component, instantaneous frequency

1. Introduction

Research on speech synthesis such as statistical parametric speech synthesis (SPSS) [1] has recently been advancing, and such synthesis requires a high-performance speech analyzer to improve the sound quality. Speech parameters (fundamental frequency (F0), spectral envelope, and aperiodicity) are widely used for SPSS. Since SPSS requires a huge amount of speech data for training, a high-performance speech analyzer would be useful not only to improve the sound quality but also to avoid having to perform post-processing by hand. There are a lot of speech analyzers to choose from these days, and the appropriate one depends on the purpose of the research. For example, real-time voice conversion [2] requires a real-time F0 estimator, whereas SPSS generally prioritizes the estimation accuracy rather than the computational cost. In this study, we focus on a high-performance F0 estimator named *Harvest* for a speech analysis/synthesis system and SPSS.

In recent SPSS, deep neural networks (DNNs) [3] utilizing continuous F0 modeling [4] have been used. This F0 modeling gives a certain F0 to the unvoiced section by an interpolation such as spline interpolation [5]. The F0 estimator preferred for this modeling should have a function that gives a smooth F0 contour to all frames. Harvest is therefore designed to reduce the error that the voiced section is wrongly identified as the unvoiced section.

In Section 2 of this paper, we discuss works related to the F0 estimation and give an overview of the proposed algorithm (Harvest). In Section 3, we explain the details of Harvest, and in Section 4, we perform an evaluation comparing Harvest with several state-of-the-art F0 estimators and discuss the results. We conclude in Section 5 with a brief summary and a mention of future works.

2. Related works on F0 estimation

F0 is defined as the shortest period of glottal vibrations. Many methods for estimating F0 have been proposed for the various purposes required. Conventional F0 estimators have used waveform features and the power spectrum [6]. Among the waveform-based algorithms, average magnitude difference function [7] and weighted auto-correlation [8] have been proposed. YIN [9] is a major estimator, and an improved version was developed [10] in 2014. As for the power-spectrum-based algorithms, methods based on cepstrum [11, 12] have been popular, and SWIPE$'$ [13] was recently proposed as a high-performance F0 estimator.

Which F0 estimator to use depends on the purpose of study. For real-time speech analysis/synthesis applications [14, 2], DIO [15] and its improved version [16] have been proposed. For high-quality speech analysis/synthesis systems, NDF [17] used in STRAIGHT [18] and XSX used in TANDEM-STRAIGHT [19, 20] are preferred. In particular, pitch synchronous analysis [21] can improve the estimation performance in the spectral envelope and aperiodicity estimation. The estimation accuracy is important in cases where the F0 is used as the input for estimating other speech parameters. CheapTrick [22, 23] used in WORLD [24], F0-adaptive multi-frame integration analysis [25], and D4C [26] require a high-performance F0 estimator. For automatic speech recognition, since the system is often used in noisy environments, a robust F0 estimator [27] would be useful.

Harvest is proposed for high-quality speech analysis/synthesis systems and for SPSS. In particular, since the continuous F0 modeling [4] gives a certain F0 to the unvoiced section, Harvest attempts to reduce the unvoiced frame and give it a reliable F0. The basic idea of Harvest is based on the event-based F0 estimator [28] and utilizes fundamental component extraction by filtering [15]. It consists of two steps: estimation of F0 candidates and generation of a reliable F0 contour on the basis of these candidates.

3. Algorithm details

We explain the details of Harvest with specific values in parameters. These values were determined after tuning to minimize the error rate in a speech database. Harvest requires a 1-ms frame shift for estimation, but users can obtain the F0 with an arbitrary frame shift by interpolation.

http://dx.doi.org/10.21437/Interspeech.2017-68

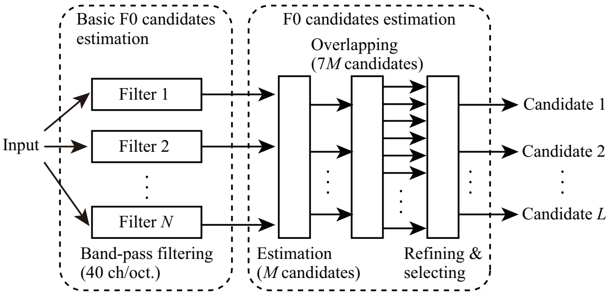
Basic F0 candidates estimation
F0 candidates estimation
Overlapping (7M candidates)
Input
Filter 1
Filter 2
Filter N
Candidate 1
Candidate 2
Candidate L
Band-pass filtering (40 ch/oct.)
Estimation (M candidates)
Refining & selecting

Figure 1: *Outline of the first step of Harvest.*

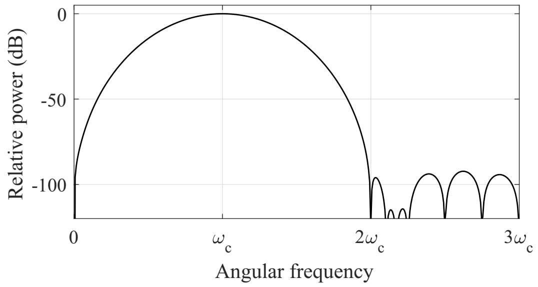

Figure 2: *Power spectrum of a band-pass filter with center frequency of ω_c Hz.*

3.1. Step 1: F0 candidate estimation

The outline of the first step is shown in Fig. 1. The purpose here is to collect all F0 candidates even if they include estimation errors. Many F0 candidates along with their reliability scores are obtained in each frame.

3.1.1. Estimation of basic F0 candidates

To begin with, the input waveform is filtered by many band-pass filters with different center frequencies. The filter $h(t)$ is designed by multiplying the Nuttall window $w(t)$ [29] and the sine wave. This is similar to the idea of YANGsaf [30].

$$h(t) = w(t)\cos(\omega_c t), \tag{1}$$

$$w(t) = 0.355768 + 0.487396\cos\left(\frac{\pi}{2T_c}t\right) + 0.144232\cos\left(\frac{\pi}{T_c}t\right) + 0.012604\cos\left(\frac{3\pi}{2T_c}t\right), \tag{2}$$

where ω_c and T_c represent the center frequency of the filter and its period ($T_c = 2\pi/\omega_c$), respectively. The range of the filter is $-2T_c < t < 2T_c$. An example of the power spectrum is shown in Fig. 2. This filter can extract the fundamental component, provided that it is included in the range near ω_c Hz. Since the F0 is unknown before estimation, many filters with different center frequencies are required. We set the center frequencies in Harvest to 40 ch/oct from the floor and ceiling frequencies.

The output signal shapes the sine wave when only the fundamental component is extracted. In this case, the four intervals shown in Fig. 3 indicate the same value. The basic F0 candidate is calculated as the inverse of their average. Harvest removes any estimated candidate that is not included in the range of $\omega_c \pm 10\%$.

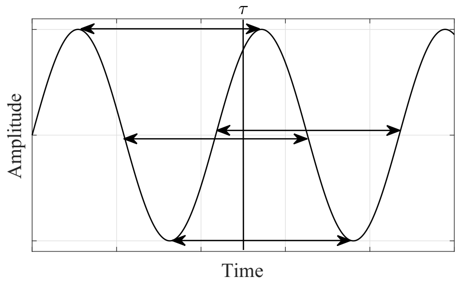

Figure 3: *Four intervals used for the estimation in a temporal position τ.*

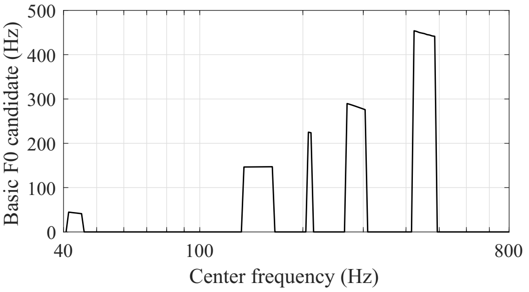

Figure 4: *Relationship between ω_c and basic F0 candidate in a frame of speech. The target F0 is around 145 Hz, but other candidates including a double-pitch error are also observed.*

3.1.2. Estimation of F0 candidates from basic F0 candidates

F0 candidates are estimated from basic F0 candidates. Figure 4 shows the relationship between the ω_c and the basic F0 candidate in a frame of speech. The horizontal and vertical axes represent ω_c and the basic F0 candidate, respectively. When the basic F0 candidate comes from the fundamental component, the same value is observed in a certain bandwidth because the filter with a center frequency near w_c outputs almost all the same waveform. Harvest obtains the F0 candidate when the filter outputs the same basic F0 candidates in a certain bandwidth. We set this bandwidth to $\omega_c \pm 10\%$ Hz.

3.1.3. Overlapping F0 candidates

Since the accuracy of this algorithm depends on the SNRs in each frame, there are often frames with no candidates due to the influence of noise. Overlapping all F0 candidates to antero-posterior frames is one way to overcome this problem. Harvest overlaps all F0 candidates to ± 3 ms.

Figure 5 shows an example of the effect of overlapping. The circle and dot represent the F0 candidate and overlapped candidate, respectively. In this example, there is no candidate in the frame of 254 ms, but the overlapping can cover this lack. All F0 candidates are refined and scored by the instantaneous frequency in the next process.

3.1.4. Refining and scoring all F0 candidates by instantaneous frequency

As we saw in Fig. 4, not only the target F0 but also several errors are estimated. To effectively select the target F0, Harvest refines and scores all F0 candidates by using the instantaneous frequency.

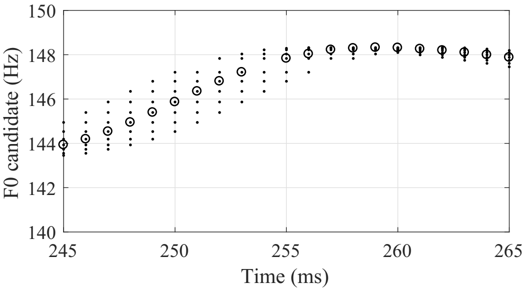

Figure 5: *Effect of overlapping. Circle and dot represent F0 candidate and overlapped candidate, respectively. There is no candidate in the frame of 254 ms, but the overlapping can cover this lack.*

Instantaneous frequency is defined as the derivative of the phase of the waveform. Flanagan's equation [31] is used to calculate the instantaneous frequency $\omega_i(\omega, t)$ by

$$\omega_i(\omega, t) \quad = \quad \frac{\Re[S(\omega,t)]\Im[\frac{\partial S(\omega,t)}{\partial t}] - \Im[S(\omega,t)]\Re[\frac{\partial S(\omega,t)}{\partial t}]}{|S(\omega,t)|^2}, \tag{3}$$

where $S(\omega, t)$ represents the spectrum of a waveform windowed by a window function shifted to t. Harvest uses a Blackman window with the width of $3T_0$, where T_0 is the inverse of the F0 candidate. $\Re[x]$ and $\Im[x]$ represent the real and imaginary parts of the input x, respectively.

The instantaneous frequency of the periodic signal indicates the value close to F0 when the frequency is around F0. Since the spectrum around F0 has a larger power, this refinement is more robust than the fundamental component extraction by the filtering. The F0 is therefore refined to a more accurate F0 even if the F0 candidate contains a certain amount of error by the noise. Actual refinement is carried out using the following equation. Refined F0 candidate $\hat{\omega}_0$ at a temporal position t is given by

$$\hat{\omega}_0 = \frac{\displaystyle\sum_{k=1}^{K} |S(k\omega_0, t)| \omega_i(k\omega_0, t)}{\displaystyle\sum_{k=1}^{K} k|S(k\omega_0, t)|}, \tag{4}$$

where ω_0 represents the angular frequency of the F0 candidate at a temporal position t, and K represents the number of harmonics used for refining. Since speech has a harmonic structure, using some harmonic components is useful for the refinement. In Harvest, we set the number of harmonics K to six.

In cases where the F0 candidate before refining equals the target F0, both the F0 candidate and the refined one indicate the same value. By the same token, $\omega_i(k\omega_0, t)$ indicates $k\omega_0$. The difference between them can therefore be used as the reliability score. This score r is given by

$$r = \frac{K}{\displaystyle\sum_{k=1}^{K} \left(\frac{\omega_i(k\omega_0, t)}{k} - \omega_0 \right)}. \tag{5}$$

In this processing, F0 candidates with scores below 2.5 are removed.

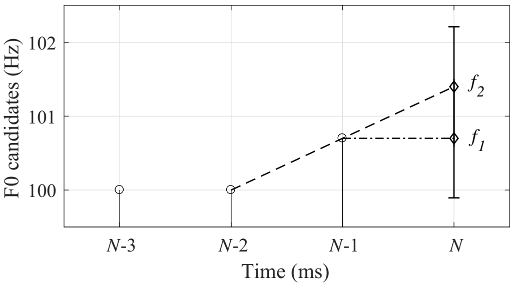

Figure 6: *Frequency range used to identify whether F0 is counted as voiced section or not.*

3.2. Step 2: Generation of the best F0 contour on the basis of estimated F0 candidates

Many F0 candidates in each frame are obtained in the first step. The purpose of the second step is to generate a reliable F0 contour from among all F0 candidates. To begin with, F0 candidates with the highest reliabilities are selected as the basic F0 contour.

3.2.1. Removal of unwanted F0 candidates

Since voiced speech is a periodic signal, the F0 contour does not change rapidly in a fundamental period on the basis of the definition of F0. F0s with a rapid change above a threshold are removed, and this frame is counted as the unvoiced section. Figure 6 shows the frequency range that can be counted as the voiced section. The frequency range in N ms is determined by F0s at $N - 1$ and $N - 2$ ms. Two frequencies f_1 and f_2 are calculated as $f_0(N - 1)$ and $2f_0(N - 1) - f_0(N - 2)$, respectively. If the F0 candidate at N ms is not included in the range of $f_1 \pm 0.8\%$ or $f_2 \pm 0.8\%$, it is removed.

3.2.2. Removal of short voiced sections

The F0 contour has at least a length based on a fundamental period, and the noise may incidentally cause a continuous F0 with a short period. Short voiced sections with a length below the threshold are removed and counted as the unvoiced section. We set the threshold to 6 ms so as to remove the unwanted section.

3.2.3. Expansion of each voiced section

Each voiced section is expanded by using the F0 candidates in unvoiced sections. An F0 candidate at voiced frame N ms is used to determine the F0 of unvoiced frame $N + 1$ ms. If the nearest F0 candidate at $N + 1$ ms is included in the range of $f_0(N)\pm 18\%$, it is selected and expanded as the voiced section. If there is no F0 candidate in $N + 1$ ms, the same process is carried out in next frame $N + 2$ ms. If there is no F0 candidate from $N + 1$ to $N + 3$ ms, the expansion process is completed. The maximum expansion is limited to 100 ms. This expansion is carried out in an anteroposterior direction. After expansion, Harvest again removes the short voiced sections. The threshold was set to 2200 / f ms, where f is the average of the F0s in the voiced section. In cases where expanded F0 contours are overlapped, the F0 contour with the higher average reliability score in the overlapped section is selected.

3.2.4. Interpolation and smoothing of the F0 contour

Since one of the aims of Harvest is to prevent the voiced section from being wrongly identified as the unvoiced section, the unvoiced sections with short periods were revised to have F0s. The unvoiced section within the period of 9 ms is regarded as the voiced section, and F0s in this section are given by the linear interpolation between the F0s of the anteroposterior voiced section of their boundaries.

The connected F0 contour is smoothed in each voiced section by a zero-lag Butterworth filter. The F0s in an unvoiced section are padded by the F0s in each boundary. After filtering, the F0 of an unvoiced section is reset to 0. The smoothing result is the final F0 contour estimated by Harvest. We set the order of the filter to two and the cut-off frequency to 30 Hz.

4. Evaluation

The proposed algorithm was evaluated using two speech databases (Japanese and English). They consist of a speech waveform and accompanying electroglottograph (EGG) signal.

4.1. Compared algorithms

We selected several state-of-the-art algorithms for the purpose of comparison: YIN[1], SWIPE$'$[2], NDF used in STRAIGHT, XSX used in TANDEM-STRAIGHT, DIO used in WORLD[3], and YANGsaf[4]. The source codes released by the authors were used to accurately evaluate each algorithm. Harvest is also released in C++[5] and Matlab. In all algorithms, the floor and ceiling frequencies of the F0 estimation range were set to 40 and 800 Hz, respectively. The frame shift was set to 1 ms. Other parameters were set to their defaults. A voiced/unvoiced (VUV) detector was used in each algorithm except for YIN, which does not have such a function.

4.2. Databases

Two speech databases were used for the evaluation. DB1 is a Japanese speech database consisting of 840 speech items uttered by 14 speakers (seven men and seven women). The parameters in Harvest were tuned using this database. Its sampling frequency was 16 kHz. DB2 is Bagshaw's speech database [32] and consists of 100 speech items uttered by two speakers (one man and one woman). Its sampling frequency was 20 kHz.

An F0 estimator is required to obtain the target F0 contour from the EGG signal. In the evaluation, the target F0 contour was estimated by NDF from the differentiated EGG signal. Although NDF can also estimate VUV information, we ignore it here and used the VUV information recorded in the databases.

4.3. Error index

Several error indices such as gross pitch error, fine pitch error [33], and gross error rate (GER) [9] have been proposed. A framework that uses artificial signals generated from an artificial F0 contour has also been proposed [34]. In cases where the EGG signal is used to obtain the target F0, we cannot discuss small differences between the target and estimated F0s because the target F0 contour depends on the performance of the F0 estimator. The simple GER was therefore used in this evaluation. Gross error is defined as a value that differs by more than 20% from the target F0. GER is defined as the ratio between the number of frames with gross error and all frames.

Table 1: *Result of evaluation by GER.*

Method	DB1	DB2
YIN	1.84	3.64
SWIPE$'$	1.19	3.99
NDF (STRAIGHT)	0.91	2.34
XSX (TANDEM-STRAIGHT)	0.97	2.32
DIO (WORLD)	1.33	4.99
YANGsaf (yang_vocoder)	0.95	5.03
Harvest	**0.33**	**1.61**

Table 2: *Ratios between the number of voiced sections in Harvest and that in others. Since YIN has no VUV detector, its result is not included.*

Method	DB1	DB2
SWIPE$'$	0.79	0.58
NDF (STRAIGHT)	0.76	0.66
XSX (TANDEM-STRAIGHT)	0.73	0.65
DIO (WORLD)	0.95	1.12
YANGsaf (yang_vocoder)	0.77	0.59
Harvest	1.0	1.0

4.4. Results and discussion

Table 1 shows the GERs for each DB. In DB1, since Harvest was tuned to minimize the GER, the performance was the best of all algorithms. Harvest also achieved the best performance of all algorithms in DB2. These results suggest that Harvest works well in both Japanese and English speech by using these tuning parameters.

In this evaluation, the VUV information recorded in each database was used and the target F0 contour depended on the estimator. By visual checking, we confirmed that the VUV information and the target F0 contour infrequently include the error. This suggests that even if an estimator can estimate the target F0, it is counted as the gross error. We therefore changed the F0 estimator for obtaining the target F0 from the EGG signal and carried out the same evaluation. Results showed that Harvest consistently achieved the best performance of all algorithms.

In this evaluation, F0 estimators with a strict VUV detector were disadvantageous. Table 2 shows the ratios between the number of voiced sections in Harvest and that in others. As shown, Harvest estimated the longest voiced section in DB1. In DB2, although DIO estimated the longest voiced section, its performance was inferior to others, which suggests that the estimated F0 was not reliable. Harvest was designed to reduce the error that the voiced section is wrongly identified as the unvoiced section, and these results show that it works as expected.

5. Conclusion

This paper described Harvest, a high-performance F0 estimator consists of the combination of fundamental component extraction and instantaneous-frequency-based refinement. Its performance was the best compared with other state-of-the-art algorithms. Our future works will include applying Harvest for speech analysis/synthesis and SPSS. Since the Merlin toolkit [35] already utilizes WORLD for vocoding speech, we can add Harvest as an F0 estimator.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16H01734, JP15H02726, JP16H05899, JP16K12511.

[1] http://audition.ens.fr/adc/sw/yin.zip

[2] http://www.cise.ufl.edu/~acamacho/english/

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

[4] https://github.com/google/yang_vocoder

[5] https://github.com/mmorise/World

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[2] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime straight speech manipulation system," *Acoust. Science & Technology*, vol. 28, no. 3, pp. 140–146, 2007.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *in Proc. ICASSP2013*, pp. 7962–7966, 2013.

[4] K. Yu and S. Young, "Continuous f0 modeling for hmm," *IEEE Trans. Audio, Speech and Language*, vol. 19, no. 5, pp. 1071–1079, 2011.

[5] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Trans. Inf. & Syst.*, vol. E97-D, no. 6, pp. 1429–1437, 2014.

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

[7] 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.

[8] 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.

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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*, pp. CD–ROM, 2009.

[16] R. Daido and Y. Hisaminato, "A fast and accurate fundamental frequency estimator using recursive moving average filters," *in Proc. INTERSPEECH 2016*, pp. 2160–2164, 2016.

[17] 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.

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[19] 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.

[20] 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.

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

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

[23] ——, "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.

[24] 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, pp. 1877–1884, 2016.

[25] 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.

[26] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[27] 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.

[28] 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.

[29] A. H. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. on acoust., speech, and signal processing*, vol. 29, no. 1, pp. 84–91, 1981.

[30] H. Kawahara, Y. Agiomyrgiannakis, and H. Zen, "Using instantaneous frequency and aperiodicity detection to estimate f0 for high-quality speech synthesis," *arXiv preprint arXiv:1605.07809*, 2016.

[31] J. Flanagan and R. Golden, "Phase vocoder," *The Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 2009.

[32] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," *in Proc. Eurospeech*, pp. 1003–1006, 1993.

[33] 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.

[34] M. Morise and H. Kawhahara, "TUSK: A framework for overviewing the performance of f0 estimators," *in Proc. INTERSPEECH 2016*, pp. 1790–1794, 2016.

[35] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *in Proc. of SSW 2016*, pp. 218–223, 2016.