HIGHER ORDER WAVEFORM SYMMETRY MEASURE AND ITS APPLICATION TO PERIODICITY DETECTORS FOR SPEECH AND SINGING WITH FINE TEMPORAL RESOLUTION

Hideki Kawahara[⋆∗] *Masanori Morise*[†] *Ryuichi Nisimura*[⋆] *Toshio Irino*[⋆]

[⋆]Wakayama University, 930 Sakaedani, Wakayama, Wakayama, 640-8510 Japan
[†]University of Yamanashi, Kofu, Yamanashi, 400-8511 Japan

ABSTRACT

Another simple and high-speed F0 extractor with high temporal resolution based on our previous proposal has been developed by adding a higher-order symmetry measure. This extension made the proposed method significantly more robust than the previous one. The proposed method is a detector of the lowest prominent sinusoidal component. It can use several F0 refinement procedures when the signal is the sum of harmonic sinusoidal components. The refinement procedure presented here is based on a stable representation of instantaneous frequency of periodic signals. The whole procedure implemented by Matlab runs faster than realtime on usual PCs for 44,100 Hz sampled sounds. Application of the proposed algorithm revealed that rapid temporal modulations in both F0 trajectory and spectral envelope exist typically in expressive voices such as those those used in lively singing performance.

Index Terms— speech analysis, fundamental frequency, speech synthesis, expressive speech, singing voices

1. INTRODUCTION

It is an interesting challenge to analyze, manipulate and resynthesize expressive speech, such as live singing performance and theatrical stage performances [1], as well as very expressive emotional speech. Such voices consist of a full range of irregularities and existing F0 extractors seem to fail capturing such irregularities. This paper revisits the very fundamental concept of F0 and proposes another simple and very fast algorithm to capture such rapid variations in the lowest prominent sinusoidal component, which usually corresponds to the fundamental component of periodic or quasi periodic signals such as voiced speech sounds. The proposed method is an extension and replacement of the method proposed in our Interspeech2012 paper [2] and outperforms it in terms of reliability. It also outperforms our proposal presented in ICASSP2012 [3].

This is a part of our ongoing project for expanding the STRAIGHT framework [4, 5, 6] for enabling analysis, modification, and resynthesis of expressive or, put another way, extreme voices. The following sections start by briefly introducing the STRAIGHT framework before describing extensions, on excitation source analyses.

2. BACKGROUND: STRAIGHT FRAMEWORK

STRAIGHT, essentially, is a source filter model. It decomposes input signal into excitation source and a sequence of spectral envelopes. The excitation source information is represented in terms of time-varying F0 and parameterized time-varying wide-band random signal. This conceptual simplicity makes STRAIGHT a powerful and easy-to-use tool for investigating human speech perception.

2.1. Interference-free power spectrum

Spectral envelope estimation of STRAIGHT consists of two stages. First stage eliminates temporal variation due to periodicity by F0-adaptive window design and F0-adaptive averaging [7]. The second stage eliminates periodicity in the frequency domain by using a rectangular smoother, the size of which is adaptively designed to match the F0 value. The consistent sampling theory [8] is used to design an F0-adaptive digital filter on the frequency axis for recovering from spectral smearing caused by an excessive smoothing effect due to time windowing and rectangular smoothing [6]. Current implementation of STRAIGHT uses cepstral littering for this F0-adaptive envelope recovery with spectral peak enhancement for improving perceptual quality of manipulated speech [9, 10]. Recently, this spectral envelope recovery [11] has been further extended on the basis of compressive sensing [12] by reformulating our previous proposal [13].

An additional important aspect of this spectral envelope estimation is in temporal resolution. It is adaptive to the fundamental period. Please note that the effective temporal resolution represented in terms of duration is slightly finer than one fundamental period [10]. This finer resolution is important for analyzing and manipulating expressiveness in singing voices [3, 2].

2.2. Excitation source analyses

In STRAIGHT framework, phase related information is intentionally discarded. Instead, only parameterized information such as fundamental frequency (F0) and aperiodicity indices are extracted. This decision is based on perception oriented design of STRAIGHT, which was originally designed to promote speech perception research, by enabling the use of ecologically relevant test stimuli [4, 5].

Current STRAIGHT uses a dense set of periodicity detectors simultaneously running and covering from 40 Hz to 800 Hz. It also has finer temporal resolution similar to that of the spectral envelope extractor of STRAIGHT [10]. This procedure successfully extracts fast temporal variations in F0 at the expense of computational complexity [1]. This computational inefficiency motivated our development of other efficient F0 extractors [3, 2].

3. SYMMETRY-BASED PERIODICITY DETECTION

This section starts by briefly discussing on related works. Voiced sounds are not always periodic. Only small regions of possible phys-

[∗]Supported by the Aid for Scientific Research No.24300073 and 24650085 from JSPS.

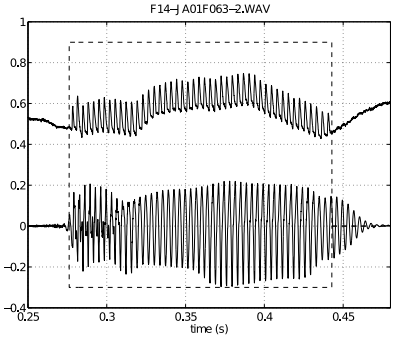

Fig. 1. EGG signal (upper plot) and speech waveform (lower plot) with manually assigned voiced part (dashed line) based on EGG. Around 0.45 s, voiced sound is generated while no GCI exists. (Excerpt from an utterance spoken by a female speaker ("keigo no tsukai kata wa muzukashii mono desu") in Japanese (in English: "It is difficult to use polite expressions properly".). This recording was done in an anechoic chamber at NAIST [24].)

ical conditions of voicing organs allow stable periodic oscillation of vocal cords [14]. Such permissible regions are surrounded by conditions that result in chaotic or multi-stable vibration pattern. Onset and offset of voicing inevitably traverse those regions and irregularities in vibration sometimes emerge during these transitions. Ordinary F0 extractors [15, 16, 17, 18], which usually assume F0 continuity or small F0 jumps, fail to analyze these irregularities properly.

Glottal closure instance (GCI) provides important information on vocal excitation and has been investigated intensively [19, 20, 21, 22, 23]. GCIs extracted by these approaches do not suffer from continuity assumption on F0. However, voicing is not always associated with closing of glottis. The vibration of vocal cord varies the area of glottis and modulates air flow even when no closure (contact of vocal cords) is actually made. Figure 1 illustrates one example of voicing without GCIs, excerpted from the database consisting of simultaneous recording of Electroglottograph (EGG) signals and speech sounds [24]. Instead of extracting GCI, we focus on modulation of air flow because it is an indicator of vocal fold vibration and has information on other sources of air flow modulation (supra-laryngeal structures) found in extreme voices such as growls [25].

3.1. Deviation measure and waveform symmetry index

Fundamental frequency f_0 is the frequency of the fundamental component. When only the fundamental component is selected, zero-crossing interval, intervals between neighboring peaks or valleys, and minimum matching shift length provide fundamental interval $T_0 = 1/f_0$. This simple operational definition has been used even in the first VOCODER [26]. Many F0 extraction methods [15] based on fundamental component selection have been proposed using band-pass filter(s) for selecting the fundamental component. Unlike those methods, ZFF [21] uses a low-pass filter for this selection (this is our interpretation).

Our proposed method also uses low-pass filters to select the fundamental component for gaining finer temporal resolution. Temporal resolution is inversely promotional to the selection filter bandwidth and is bounded by the time-frequency uncertainty limit. Using low-pass filters to select the fundamental component makes bandwidth of the selection filter twice as wider as bandpass filters designed for selecting the fundamental component only and consequently allows twice the finer temporal resolution. This is because low-pass filters

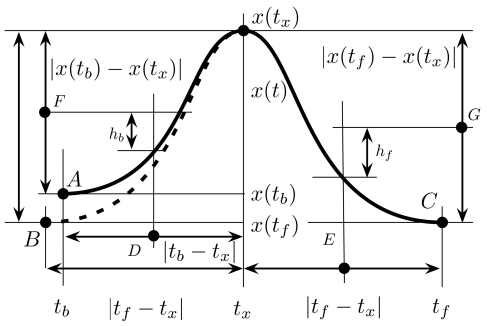

Fig. 2. Definition of reference points for measuring deviation from symmetry. Dashed line shows the mirror image of the latter half cycle. Temporal position t_x represents the maximum position, and t_b and t_f represent the temporal positions of the preceding and the following extrema.

cover zero, negative and positive frequency components while the band-pass filter has to separate each component.

Selection of the fundamental component requires information about the fundamental frequency. However, it is generally not available in advance. The proposed method uses a set of low-pass filters covering from 32 Hz to 1000 Hz with six filters in each octave and selects the best filter on the basis of a deviation measure of waveform symmetry of each filter output.

Figure 2 shows reference points to define the deviation measure. Points A and C represent neighboring extrema of a maximum point at time t_x. When the waveform is a sinusoid, temporal mirror image of C (represented as B in the figure) overlaps with A and $T_0 = |t_f - t_b|$ represents the fundamental period. In other words, the distance between A and B, deviation from symmetry, represents the amount of deviation from sinusoid [2].

In addition to these first order reference points, four reference points (D, E, F and G) are introduced to define the higher order asymmetry, this time, vertical waveform asymmetry, for each half cycle. The middle temporal positions of each half cycle D and E are used to measure the waveform deviations (h_b and h_f) from the level (vertical) middle points of each half cycle F and G.

These reference points provide three deviation values for each extreme point (each half pitch period), d_{AM}, d_{FM}, and d_{SM} (A: amplitude, F: frequency, and S: symmetry), which are normalized by the amplitude, T_0, and the amplitude of each half cycle, respectively. The symmetry index $\eta_E(x, k, f)$ for representing "goodness" of the waveform symmetry is derived from these deviation values.

$$\eta_E(x,k,f) = \exp\left(-\alpha\left(\sum_{q\in K} w_q \tilde{d}_q^{\beta}(x,k,f)\right)^{\frac{1}{\beta}}\right) \qquad (1)$$

$$\sum_{q\in K} w_q = 1\,, \quad \tilde{d}_q = \frac{d_q[k-1] + d_q[k] + d_q[k+1]}{3\sqrt{V(d_q)}},$$

where $K = \{AM, FM, SM\}$ represents the set of deviation category, and k represents the index of the extreme point, and f represents the nominal frequency of the low-pass filter. Parameter α is introduced to shape the symmetry index to have values inside $[0, 1]$, where the index value 1 represents perfect symmetry. The weights $w_q, (q \in K)$ and Minkowski distance parameter β are introduced to combine three deviation values properly: d_{AM}, d_{FM}, and d_{SM}. The variance $V(d_q)$ is the variance of d_q for Gaussian random signal input to the low-pass filter. A set of simulation tests was conducted

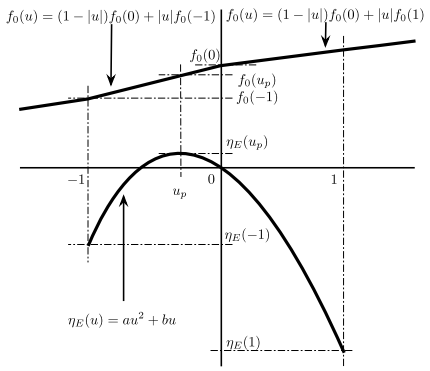

Fig. 3. Initial F0 value of candidates are linearly interpolated using the interpolated maximum location information obtained by parabolic interpolation around the maximum discrete point. In this figure the filter index u is set to 0 for a local maximum and are set to -1 and 1 for neighboring filter channels. The interpolated index u_p is used to calculate the linearly interpolated frequency $f_0(u_p)$.

to select these parameters. Details of the tests are given in the following sections.

Preliminary study indicated that impulse response of the low-pass filters have to be temporally bounded and have to be positive definite. Its frequency response has to have very low side lobe levels (lower than -80 dB) and fast decay (steeper than -12 dB/oct). Taking these conditions into account, we decided to use time windowing functions as the impulse response of low-pass filters. However, usual windowing functions [27] do not meet these requirements. One practical selection is to use one of Nuttall windows reported in the literature [28] with four cosine terms and sidelobe decay rate of -18 dB/oct. (Note that the selected window is not the commonly known "Nuttall window." Coefficients for zeroth through third cosine of the selected window are 0.355768, 0.487396, 0.144232, and 0.012604, respectively. Refer to item 12 of Table II in the work of Nuttall [28].)

3.2. Fundamental component selection and F_0 measurement

For each extreme point of filtered waveform, extrema of the symmetry indices $\eta_E(x,k,f), f \in \{f_1, f_2, \ldots, f_N\}$ are extracted, where f_n represent the nominal frequency of the n-th low-pass filter. The frequency of each candidate of the fundamental component is calculated using parabolic interpolation and linear interpolation as shown in Fig. 3. For each filter output, the candidate frequency $f_0 = 1/T_0$ is calculated from the interval between neighboring maxima or minima, T_0. In the initial estimate, the candidate that has the highest symmetry index $\eta_E(x,k,f)$ is selected as the fundamental component, and the corresponding interpolated frequency provides the initial estimate of the fundamental frequency.

3.3. Parameter tuning

A set of simulation tests was conducted to tune parameters involved in the proposed method. Let define a set consisting of the target parameters: $\Theta = [\{w_q\}_{q\in K}, \alpha, \beta]$. The cost function for this tuning is the error probability $P_E(\Theta, r)$ of selecting the fundamental component in given signal to noise ratio (SNR) conditions.

$$P_E(\Theta, r) = P\left(P_N(\eta_E > \theta) = P_S(\eta_E < \theta) | \Theta, SNR = r\right), \quad (2)$$

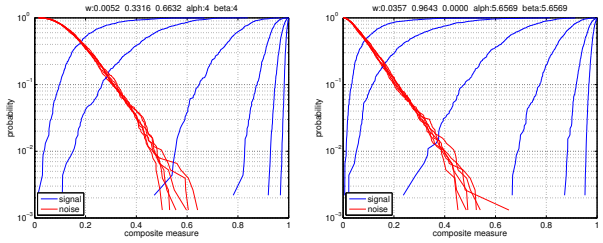

Fig. 4. Probability of symmetry index values using the best parameter set for this proposal $\Theta_{3w}(20\text{dB})$ (left plot) and the best parameter set for our previous proposal [2] $\Theta_{2w}(20\text{dB})$ (right plot). The red lines represent the probabilities for Gaussian noise input. The blue lines represent the probabilities for periodic pulse plus noise with different SNR settings. SNR for the blue lines are 0 dB, 10 dB, 20 dB, 30 dB, 40 dB, and 50 dB, from left to right.

where r represents the SNR. $P_N(\eta_E > \theta)$ represents the probability of the symmetry index η_E of the best candidate to exceed the given threshold value θ for Gaussian random noise inputs. $P_S(\eta_E < \theta)$ represents the probability of the symmetry index η_E of the best candidate to have a value smaller than the given threshold value θ for multiple sinusoids plus noise inputs. The initial phase of each component is randomized and tested 10 times for each SNR condition.

Let us define the best parameter set $\Theta_{3w}(r)$, where all three weights $\{w_q\}_{q\in K}$ are used.

$$\Theta_{3w}(r) = \arg\min_{\Theta} P_E(\Theta, r). \quad (3)$$

By using this formulation, our Interspeech2012 proposal [2] is represented as the special case, where the constraint $w_{SM} = 0$ is imposed. The best parameter set $\Theta_{2w}(r)$ for this case is defined as follows.

$$\Theta_{2w}(r) = \arg\min_{\Theta} \big|_{w_{SM}=0} P_E(\Theta, r). \quad (4)$$

The simulation results for $\Theta_{3w}(r)$ indicated that the weight for the AM deviation is significantly smaller than weights for FM and SM (0.0052, 0.3316 and 0.6632 for AM, FM, and SM respectively). Also, the shaping parameter $\alpha = 4$ and the Minkowski exponent $\beta = 4$ found to yield the best performance.

Figure 4 shows probabilities of the highest symmetry index for parameter sets $\Theta_{3w}(20\text{dB})$ and $\Theta_{2w}(20\text{dB})$. These plots indicate that introduction of a higher deviation measure significantly reduces the error probability $P_E(20\text{dB})$ from 15% (for $\Theta_{2w}(20\text{dB})$) to 5% (for $\Theta_{3w}(20\text{dB})$).

4. INSTANTANEOUS FREQUENCY-BASED REFINEMENT

The proposed method only uses the lowest (locally close to) sinusoidal component and is not noise tolerant. When the target signal is a periodic signal and consists of other harmonic components than the fundamental, instantaneous frequencies of other harmonic components can be used to refine the initial estimate, which is calculated only from the fundamental component.

A two-stage refinement procedure is introduced using a stable representation of instantaneous frequency by using two time windows [29]. In the first stage, the first and second harmonic components are utilized for refinement. This restriction is introduced to make F0 errors within $\pm 20\%$ recoverable. In the second stage, the first to the sixth harmonic components are used. This refinement procedure is the same as that in our previous proposal [2] and was

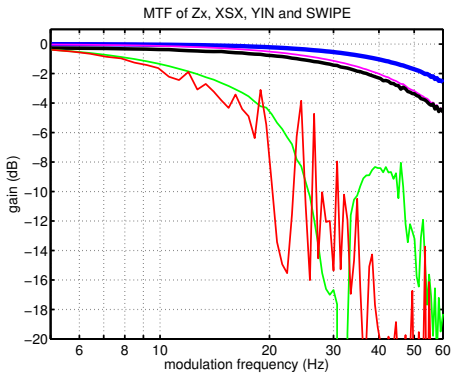

Fig. 5. Modulation transfer function for F0 frequency modulation: (blue) initial estimates of the proposed method, (violet) refined estimates using instantaneous frequencies, (black) XSX, (green) YIN and (red) SWIPE′. These results are the same as those of our previous proposal [2].

found to reduce the fine estimation error one-tenth for SNR ranging from 20 dB to 50 dB.

5. NUMERICAL RESULTS

This section illustrates numerical aspects of the proposed method using artificial test signals and natural voices. For the processing speed of the proposed method, the initial estimation stage requires 0.6 s to process 3.3 s speech (44,100 Hz sampling, 1 ms frame rate), and the refinement stage requires 2.2 s for the same speech (OS X 10.8.2, PowerBook Pro 2.6 GHz Intel Core i7 16GB memory and MATLAB R2012a).

5.1. Tracking to frequency modulation

Figure 5 shows the modulation transfer function of a FM test signal with multiple harmonic components. Modulation frequency of the frequency modulation is started from 4 Hz and log-linearly increased to 64 Hz. The carrier frequency was 200 Hz. The plot shows results using the proposed method, F0 extractor of the current TANDEM-STRAIGHT (XSX), YIN [16], and SWIPE′ [17]. The last two methods were selected as commonly used references. This plot clearly indicates that the proposed F0 extractors including XSX can track very fast frequency modulation of the fundamental component, while existing F0 detectors fail. Note that these results are identical to those of our previous proposal [2] since the proposed method improves only reliability of the initial candidate selection process. For further improving temporal resolution, the Teager-Kaiser Energy Operator [30, 31] can be used, instead of using interval measurement of T_0, at the expense of noise sensitivities.

5.2. Tracking performance for natural speech inputs

Tracking performance for natural speech inputs was tested using two speech databases with simultaneous EGG recordings [24, 32]. The database recorded at NAIST [24] consists of 30 Japanese sentences each for 28 (14 male and 14 female) speakers. The database by Bagshaw [32] consists of 50 sentences each spoken by one male and one female speaker. The performance of the initial estimation stage of the proposed method represented in terms of the Pitch Tracking Error (PTE), proposed by Ellis [18], was 6% for the NAIST database

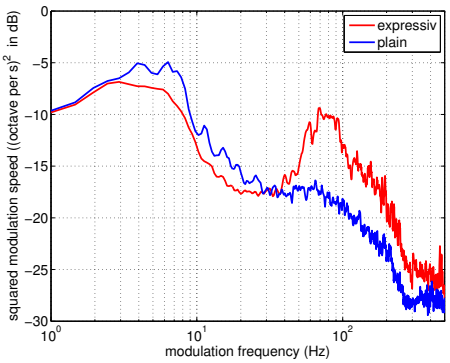

Fig. 6. Power spectra of differentiated logarithmic fundamental frequency trajectories calculated for normal performance (blue line) and expressive performance (red line) of a Japanese POP song (“Ride”) sang by a male professional singer.

and 7% for the Bagshaw database in clean speech condition. This performance is comparable with that of the SAcC method [18] and outperforms YIN and SWIPE′. Note that the ground truth values of voicing in these evaluation results are based on EGG signals and cannot be very dependable because of the issues illustrated in Fig. 1.

5.3. Expressive singing

Figure 6 shows the power spectra of the differentiated logarithmic fundamental frequency trajectories extracted by the proposed method with refinement, from singing voices in two different expressions. The song is the original J-POP song “Ride”, composed for research use. The song is 157 s long, and only voiced segments were analyzed. It was performed by a Japanese professional singer in his usual expressive singing (full of “shout”) and in plain singing style. The figure illustrates that the expressive singing consists of strong modulation energy around 70 Hz. In some of the very expressive passages, the peak magnitude of modulation was over 20 dB higher than that of its plain counterpart. It is also observed that the STRAIGHT spectrogram of this expressive singing shows temporal modulations synchronized with F0 modulations. Preliminary manipulation and resynthesis tests using STRAIGHT suggested that these observed features are contributing factors for our perception of “shout” expression in singing.

6. CONCLUSIONS

Another simple and high-speed F0 extractor with high temporal resolution was developed by introducing higher-order symmetry measure to our previous symmetry-based method. Temporal fine structures found in both excitation source and spectral envelope by applying the proposed method seem to play important roles in expressive or extreme voices. Perceptual effects of these structures and integration of this method and other acoustic event related procedures to STRAIGHT framework are currently being systematically investigated.

7. ACKNOWLEDGEMENTS

The authors thank B. Yegnanarayana for inspiring discussions on event detection and zero frequency filtering. They also thank O. Fujimura and K. Sakakibara for discussions on vocal fold physiology and their aperiodic behavior.

8. REFERENCES

[1] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J.C. Williams, "Noh voice quality," *J. Logopedics Phoniatrics Vocology*, vol. 34, no. 4, pp. 157–170, 2009.

[2] M. Kawahara, H. Morise, R. Nisimura, and T. Irino, "Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulatio," in *Interspeech2012*, 2012, Session: O2d.05.

[3] H. Kawahara, M. Morise, and T. Irino, "Analysis and synthesis of strong vocal expressions: Extension and application of audio texture features to singing voice," in *ICASSP2012*, 2012, pp. 5389–5392.

[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 Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[5] H. Kawahara, "STRAIGHT, exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustic Science & Technology*, vol. 27, no. 5, pp. 349–353, 2006.

[6] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," *ICASSP2008*, pp. 3933–3936, 2008.

[7] M. Morise, T. Takahashi, H. Kawahara, and T. Irino, "Power spectrum estimation method for periodic signals virtually irrespective to time window position," *Trans. IEICE*, vol. J90-D, no. 12, pp. 3265–3267, 2007, [in Japanese].

[8] M. Unser, "Sampling – 50 years after Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, 2000.

[9] H. Akagiri, M. Morise, T. Irino, and H. Kawahara, "Evaluation and optimization of F0-adaptive spectral envelope estimation based on spectral smoothing with peak emphasis," *Trans. IEICE*, vol. J94-A, no. 8, pp. 557–567, 2011, [in Japanese].

[10] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *SADHANA*, vol. 36, no. 5, pp. 713–722, 2011.

[11] H. Kawahara, T. Toda, R. Nisimura, and T. Irino, "Beyond bandlimited sampling of speech spectral envelope imposed by the harmonic structure of voiced sounds," in *ICASSP2013*, 2013, (submitted).

[12] Yonina C. Elder and Tomer Michaeli, "Beyond bandlimited sampling," *Signal Processing Magazine*, pp. 48–68, May 2009.

[13] H. Kawahara, M. Morise, R. Nisimura, and T. Irino, "Spectral envelope recovery beyond the Nyquist limit for high-quality manipulation of speech sounds," in *Interspeech2008*, 2008, pp. 650–653.

[14] Ingo R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2733–2749, May 2008.

[15] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag, 1983.

[16] A. de Chevengné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[17] 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.

[18] B. S. LEE and D. P. W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Interspeech2012*, 2012, Session: P3b.O5.

[19] S. K. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 917—924, 1992.

[20] H. Kawahara, Y. Atake, and P. Zolfaghari, "Accurate vocal event detection method based on a fixed-point to weighted average group delay," in *ICSLP2000*, 2000, pp. 664–667.

[21] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[22] N. Sturmel, C. d'Alessandro, and F. Rigaud, "Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform," in *ICASSP2009*, 2009, pp. 4517—4520.

[23] Mark R. P. Thomas, Jon Gudnason, and Patrick A. Nay, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio Speech and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[24] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," in *Proc. ICSLP*, Beijin, 2000, pp. 907–910.

[25] K. Sakakibara, H. Fuks, N. Imagawa, and N. Tayama, "Growl voice in ethnic and Pop styles," in *Proc. Int. Symp. on Musical Acoustics*, 2004.

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

[27] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[28] A. H. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. Audio Speech and Signal Processing*, vol. 29, no. 1, pp. 84–91, 1981.

[29] H. Kawahara, T. Irino, and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," *ICASSP2011*, pp. 5420 –5423, may 2011.

[30] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "On amplitude and frequency demodulation using energy operators," *Signal Processing, IEEE Transactions on*, vol. 41, no. 4, pp. 1532 –1550, apr 1993.

[31] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *Signal Processing, IEEE Transactions on*, vol. 41, no. 10, pp. 3024 –3051, oct 1993.

[32] P. Bagshaw, *Automatic prosodic analysis for computer aided pronunciation teaching*, Phd thesis, University of Edinburgh, 1994.