ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulation

Hideki Kawahara[1], *Masanori Morise*[2], *Ryuichi Nisimura*[1], *Toshio Irino*[1] *

[1]Faculty of Systems Engineering, Wakayama Univ., Wakayama, Wakayama, 640-8510 Japan
[2]College of Information Science and Eng., Ristumeikan Univ., Kusatsu, Shiga, 525-8577 Japan
{kawahara,nisimura,irino}@sys.wakayama-u.ac.jp, morise@fc.ritsumei.ac.jp

Abstract

A simple and high-speed F0 extractor with high temporal resolution is proposed based on a waveform symmetry measure. Strictly speaking, it is not an F0 extractor. Instead, it is a detector of the lowest prominent sinusoidal component with a salience measure. It can make use of an F0 refinement procedure, when the signal under investigation is a sum of harmonic sinusoidal components. The refinement procedure is based on a stable representation of instantaneous frequency of periodic signals. Application of the proposed algorithm revealed that rapid temporal modulations in both F0 trajectory and spectral envelope exist typically in expressive voices such as lively singing performance. Manipulation of these temporal fine structures (texture) effectively modified perceptual expressiveness, while somewhat preserving perceptual vocal effort and register.

Index Terms: speech analysis, speech synthesis, expressive speech, singing voices

10.21437/Interspeech.2012-139

1. Introduction

It is an interesting challenge to analyze, manipulate and resynthesize expressive speech, such as live singing and theatrical stage performances, and very expressive emotional speech. The vocal elements of such speech typically contain of a full range of irregularities, and conventional F0 extractors are unable to capture these irregularities. In this paper, we revisit the fundamental concept of F0 and propose a simple, very fast algorithm to capture 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 a complete reformulation and replacement of the method proposed in our ICASSP2012 paper [1] and outperforms it in terms of speed and accuracy.

This study is a part of our ongoing extension of the STRAIGHT framework [2, 3, 4] for the analysis, modification and resynthesis of expressive (i.e., extreme) voices. Here, we give a brief overview of the STRAIGHT framework followed by a description of our extensions, with a focus on those related to excitation source analyses.

2. Background: STRAIGHT framework

STRAIGHT is essentially a source filter model. It decomposes an input signal into an excitation source and a sequence of spectral envelopes. The excitation source information is represented by a time-varying F0 and parameterized time-varying wide-band random signal. This conceptually simple decomposition makes STRAIGHT a powerful tool for investigating human speech perception.

* Partly supported by Grants-in-Aid for Scientific Research 22650042 and 23700221 by JSPS, Advanced Research Initiative by Wakayama University and Ono Acoustics Research Fund.

2.1. Interference-free power spectrum

The spectral envelope estimation performed by STRAIGHT consists of two stages: first, the elimination of temporal variations due to periodicity by F0-adaptive window design and F0-adaptive averaging [5], and second, the elimination of periodicity in the frequency domain by using a rectangular smoother, the size of which is adaptively designed to match the F0 value. We adopted the consistent sampling theory [6] to design an F0-adaptive digital filter on the frequency axis for recovering from spectral smearing caused by excessive smoothing effect due to time windowing and rectangular smoothing [4].

Recently, a new implementation of this procedure based on the logarithmic conversion of spectral information has been introduced to improve recovery accuracy and perceptual sound quality [7, 8]. In this implementation, spectral envelope $P_{ST}(\omega)$ is calculated from the temporally stable power spectrum $P_T(\omega)$ by

$$P_{ST}(\omega) = \exp\left(\mathcal{F}\left[g_1(q)g_2(q)C_T(q)\right]\right), \quad (1)$$

where $C_T(q)$ represents the cepstrum of $P_T(\omega)$ and q represents quefrency. Symbol $\mathcal{F}[\,]$ represents the Fourier transform. Two lifters, $g_1(q)$ and $g_2(q)$, are defined as

$$g_1(q) = \tilde{\alpha}_0 + 2\tilde{\alpha}_1 \cos\left(2\pi q f_0\right), \quad (2)$$

$$g_2(q) = \frac{\sin(\pi f_0 q)}{\pi f_0 q}, \quad (3)$$

where $f_0 = 1/T_0$ represents the fundamental frequency (F0). The second lifter ($g_2(q)$) represents the F0-adaptive spectral smoothing using the rectangular smoothing function (width is set to f_0), and the first lifter ($g_1(q)$) represents a digital filter on the frequency axis that compensates for over-smoothing due to $g_2(q)$ and the time windowing used to calculate $P_T(\omega)$. A detailed discussion of this cepstrum-based implementation has been provided in a previous work [7].

The filter coefficients $\tilde{\alpha}_0$ and $\tilde{\alpha}_1$ of $g_1(q)$ are numerically determined to minimize the Itakura-Saito spectral distance on the perceptual frequency axis (ERB_N number [9]) for various types of vocal tract shapes [10] and excitation source variations using the LF-model [11]. Subjective tests revealed that this new implementation using the optimized coefficients $\tilde{\alpha}_0 = 1.18, \tilde{\alpha}_1 = -0.09$ provides better manipulated speech sounds [8] than both previous STRAIGHT implementations [2, 4] and PSOLA [12]. Since the final form of this implementation is Cepstrum liftering, we felt it would be beneficial to investigate its relation to the latest True-Envelope based approach [13].

An important advantage of our spectral envelope estimation is its temporal resolution, which 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 [7]. This finer resolution is crucial for analyzing and manipulating expressiveness in speech sounds.

2.2. Excitation source estimation

In the STRAIGHT framework, phase information is intentionally discarded and only parameterized information such as fundamental frequency (F0) and aperiodicity indices is extracted. This decision is based on the perception oriented design of STRAIGHT, which was originally intended to promote speech perception research, by providing means to use only ecologically relevant test stimuli [2, 3]. Extracted source information is used to generate a pulse plus noise excitation source signal in the synthesis stage of the framework.

The currently used STRAIGHT excitation information extractor (called eXcitation Structure eXtractor, or XSX) uses a dense set of periodicity detectors (e.g., four detectors in one octave) that simultaneously run and cover a range of 40 Hz to 800 Hz. It also has a finer temporal resolution, similar to its spectral envelope extractor [7]. This procedure can be used to successfully extract fast temporal variations in F0, though at the expense of computational complexity [14]. This computational inefficiency is what motivated us to develop a new F0 extractor.

3. Symmetry-based periodicity detection

Voiced sounds are not always periodic, in fact, only small regions of the possible physical conditions of voicing organs permit stable periodic oscillation of the vocal cords [15]. Such permissible regions are surrounded by various conditions, which results in a chaotic or multi-stable vibration pattern. The onset and offset of voicing inevitably traverse those regions, and irregularities in vibration sometimes emerge. F0 extractors, 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 extensively [16]. Among the various GCI estimation methods, zero frequency filtering (ZFF) delivers the best performance [16]. However, voicing does not always require glottal closing: vocal cord vibration varies the glottal area and modulates air flow even when no closure (contact of vocal cords) is actually made.

We focus on the modulation of air flow rather than the extraction of GCI because air flow is an indicator of vocal fold vibration and contains information on other sources of air flow modulation found in extreme vocal sounds such as growl [17]. The proposed method, while inspired by ZFF, is based on a different interpretation of its underlying principle, as we discuss below.

3.1. Deviation measure of waveform symmetry

Fundamental frequency f_0 is the frequency of the fundamental component. When only the fundamental component is selected, the 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 in use ever since the first VOCODER was developed [18]. Many F0 extraction methods based on fundamental component selection have been proposed using band-pass filter(s) for selecting the fundamental component. Unlike those methods, ZFF essentially uses a low-pass filter for this selection (in our interpretation, low-pass filtering is essential and zero frequency is not).

Our proposed method also uses low-pass filters to select the fundamental component for gaining finer temporal resolution. Temporal resolution is inversely proportional to the selection filter bandwidth and is bounded by the time-frequency uncertainty limit. Using low-pass filters to select the fundamental component makes the bandwidth of the selection filter two times wider than the bandpass filters and consequently enables a temporal resolution two times finer. This is because low-pass filters can cover zero, negative and positive frequency components while band-pass filters have to separate each component.

We need information about the fundamental frequency in order to select the fundamental component, but this information is generally not available in advance. The proposed method uses a set of low-pass filters covering a range of 32 Hz to 1000 Hz with 6 filters in each octave and selects the optimum filter on the basis of deviation measurements of the wave symmetry of each filter output.

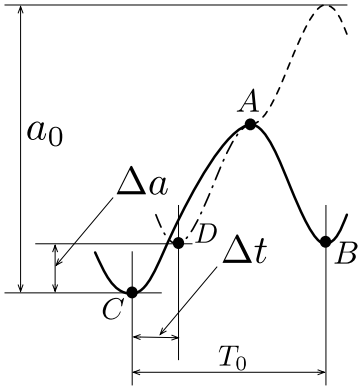

Figure 1: Definition of relative deviation from symmetry. Dashed line is the amplitude mirror image of the latter half of the cycle. Dash-dot line is the temporal mirror image of the latter half of the cycle.

Figure 1 shows reference points to define the deviation measure. Points A, B and C represent the neighboring extrema of a waveform. When the waveform is a sinusoid, the temporal mirror image of B (D in the figure) overlaps with C, and T_0 represents the fundamental period. In other words, the distance between C and D, which is the deviation from symmetry, represents the amount of deviation from the sinusoid. This enables us to evaluate the relative magnitude of the first order coefficients of the Taylor expansion of modulations $b(t)$ and $f(t)$ (amplitude and frequency, respectively):

$$y(t) = b(t) \sin\left(2\pi \int_0^\tau f(\tau) d\tau + \varphi\right), \tag{4}$$

where $y(t)$ represents a modulated sinusoid and φ represents the initial phase. Note that, the higher order coefficients of the Taylor expansion can also be taken into account by adding other reference points for testing symmetry.

The final form of the periodicity index derived from deviation measure is

$$\eta_E = \exp\left(-\alpha \left(\frac{(\beta \Delta t)^\gamma}{T_0^\gamma} + \frac{\Delta a^\gamma}{a_0^\gamma}\right)^{1/\gamma}\right), \tag{5}$$

where β represents the tuning coefficient for equating effects of amplitude modulation and frequency modulation and α represents the tuning coefficient for shaping the level distribution of η_E for random signals. The exponent γ for the Minkowski distance is used to fit the distance to the magnitude of deviation from sinusoids. The tuning coefficients and the exponent values are optimized for a 20 dB SNR condition using simulations ($\alpha = 4.5$, $\beta = 2.6$, and $\gamma = 9$). For sinusoids, the periodicity index indicates 1 and decreases when the deviation measure increases.

In the present work, preliminary study indicated that the impulse response of the low-pass filters must be temporally bounded and positive definite. Their frequency responses must have very low side lobe levels (under -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 the low-pass filters. However, standard windowing functions [19] do not meet our requirements. We therefore opted for one of the Nuttall windows reported in the literature [20] with four cosine terms and sidelobe decay rate of

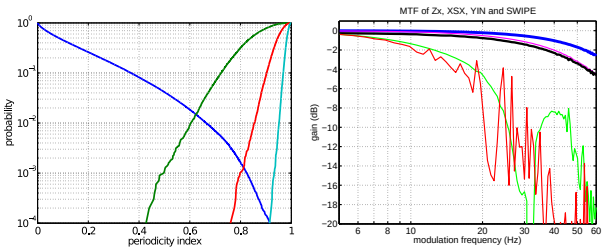

Figure 2: [left plot] Distribution of peak periodicity index for pulse train with different SNR (cyan line: 40 dB, red line: 30 dB and green line: 20 dB) and Gaussian white noise (blue). [right plot] Modulation transfer function for F0 frequency modulation. (blue) proposed method without refinement, (violet) proposed method with refinement, (black) XSX, (green) YIN and (red) SWIPE'.

-18 dB/oct. (Note that the window we used is not the commonly known "Nuttall window." The coefficients for zero through third cosines are 0.355768, 0.487396, 0.144232 and 0.012604, respectively. Refer to item 12 of Table II in the literature [20].)

4. Instantaneous frequency-based refinement

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

A two stage procedure is introduced to refine the estimated F0. Both stages use a temporally stable representation of instantaneous frequency [21]. In the first stage, components up to the second harmonic are utilized. This restriction is introduced to ensure that F0 errors within $\pm 20\%$ are recoverable. In the second stage, components up to the sixth harmonic are used.

5. Numerical examples

In this section, we describe the numerical aspects of the proposed method using artificial test signals and natural voices. Test results showed that the proposed method without refinement requires 0.2 s to process 1 s of speech (44,100 Hz sampling, 1 ms frame rate) and that the refinement stage requires 0.5 s for the same speech (OS X 10.7.3, PowerBook Pro 2.8 GHz Intel Core i7 8GB and MATLAB R2011a).

5.1. Response to artificial test signals

The first test signal we used was a periodic pulse train plus Gaussian white noise with different SNR. The input pulse frequency was 100 Hz. The left plot of Fig. 2 shows the distribution of peak values for white noise input and the test signals with different SNR. The horizontal axis is a threshold ζ and the vertical axis is $P_r(\eta_E > \zeta)$ for noise input and $P_r(\eta_E < \zeta)$ for test signals. Setting the voicing decision threshold to 0.8 yielded a gross error rate of 0.15% when SNR was higher than 30 dB. Pulse plus noise testing showed that the refinement procedure made the F0 estimation rms errors 1/10 when SNR is higher than 20 dB. At 20 dB SNR, the rms error for F0 value was 0.1% when the refinement procedure was applied.

In the next test, we used a harmonic sinusoid generated from a frequency modulated fundamental component to test the modulation transfer functions of the F0 extractors. Modulation frequency of the frequency modulation was started from 4 Hz and log-linearly increased to 64 Hz. The carrier frequency was 200 Hz. The right plot of Fig. 2 shows the results using the proposed method, XSX, YIN [22] and SWIPE' [23]. These latter two methods were selected as reference due to their common use. The plot clearly indicates that the proposed F0 extractors, including XSX, are capable of tracking very fast frequency modulations of the fundamental component, while the conventional F0 detectors fail.

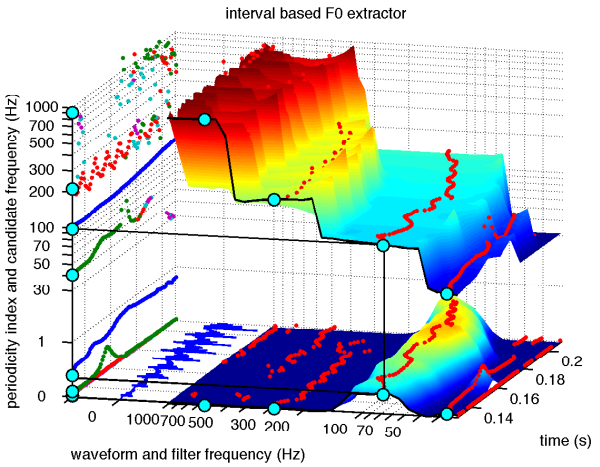

Figure 3: Snapshot of movie visualization of proposed method.

5.2. Natural speech (voice)

Figure 3 shows a snap shot of a movie visualization of the proposed method–specifically, part of our analysis of a Japanese vowel sequence (/aiueo/) spoken by a male. The colored surface on the bottom is a 3D representation of the periodicity index, while the floating colored surface is a 3D representation of the repetition frequency calculated from the interval measurement mentioned above. The red dots on the lower surface indicate the maxima on each frame. (Location of maxima is calculated using parabolic interpolation around the peak value.) These locations were projected onto the floating surface and used to read the F0 candidate frequency. (F0 candidate frequency was calculated using linear interpolation around the projected points.) In this visualization these maxima and the F0 candidates were projected onto the left wall and the input waveform was also displayed on the floor. The original movie is in the attached media files and is also available on the Web [24].

We also used the proposed method to analyze two singing samples sang by a male pop singer. The singer was asked to sing an original Japanese pop song in two different styles: his normal expressive performance and a plain (less expressive) performance. The lower left plot of Fig. 4 shows an expanded view around a major F0 transition around 400 ms. The lower right plot of Fig. 4 shows the power spectra of differentiated F0 trajectories. The figure clearly indicates that the expressive performance consists of a significantly higher (about 20 dB) modulation power around 70 Hz, which conventionall F0 extractors cannot investigate. Interestingly, a spectrographic representation of the spectral envelope sequence extracted by STRAIGHT using these results also revealed a temporal fine texture.

6. Discussion

We performed a preliminary analysis, modification and synthesis test in order to investigate the effect of these fine temporal structures. A simple resynthesis procedure replicated the original singing almost perfectly. Smoothing these structures by moving averaging using a 20 ms Hann window as the smoother seemed to selectively eliminate the liveliness of the expression while preserving the voice register and effort. Specifically, the expressive singing voice, categorized as a "musical growl" was transformed into a more typical "strong and loud" voice. These sound samples are available on the Web [24]. Although this is an informal observation, it may have interesting implications

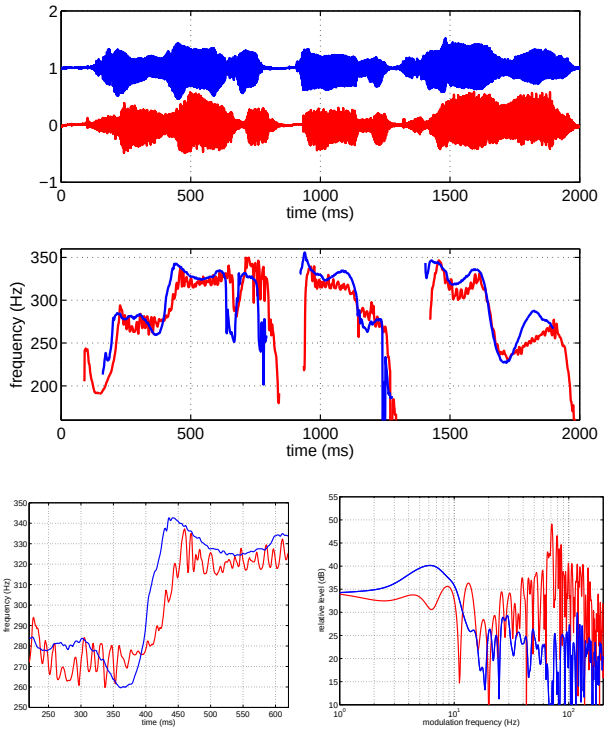

Figure 4: [Top plot] Waveform of singing performance. [middle plot] F0 trajectories of the performance. [bottom left plot] Expanded F0 trajectories around musical note transition. [bottom right plot] Power spectra of differentiated F0 trajectories. (red) expressive performance and (blue) plain performance.

for future research on expressive voices.

7. Conclusions

We proposed a simple and high-speed F0 extractor with high temporal resolution based on a waveform symmetry measurements. Temporal fine structures found in both the excitation source and the spectral envelope by using the proposed method seem to play important roles in expressive or extreme voices. We are currently working on systematic investigations into the perceptual effects of these structures and on integrating the proposed method and other acoustic event-related procedures into the STRAIGHT framework.

8. 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.

9. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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].

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

[7] 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.

[8] 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].

[9] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[10] B. H. Story, "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," *J. Acoust. Soc. Am.*, vol. 26, no. 1, pp. 327–335, 2008.

[11] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 505–519, 1995.

[12] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453 – 467, 1990.

[13] V. Villavicencio, A. R obel, and X. Rodet, "Applying improved spectral modeling for high quality voice conversion," *ICASSP2009*, pp. 4285–4288, 2009.

[14] 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.

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

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] ," http://www.wakayama-u.ac.jp/ kawahara/Interspeech2012/.