9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis

Hideki Kawahara[1,2], *Yannis Agiomyrgiannakis*[1], *Heiga Zen*[1]

[1]Google
[2]Wakayama University, Japan
kawahara@sys.wakayama-u.ac.jp,{agios,heigazen}@google.com

Abstract

This paper introduces a general and flexible framework for F0 and aperiodicity (additive non periodic component) analysis, specifically intended for high-quality speech synthesis and modification applications. The proposed framework consists of three subsystems: instantaneous frequency estimator and initial aperiodicity detector, F0 trajectory tracker, and F0 refinement and aperiodicity extractor. A preliminary implementation of the proposed framework substantially outperformed (by a factor of 10 in terms of RMS F0 estimation error) existing F0 extractors in tracking ability of temporally varying F0 trajectories. The front end aperiodicity detector consists of a complex-valued wavelet analysis filter with a highly selective temporal and spectral envelope. This front end aperiodicity detector uses a new measure that quantifies the deviation from periodicity. The measure is less sensitive to slow FM and AM and closely correlates with the signal to noise ratio. The front end combines instantaneous frequency information over a set of filter outputs using the measure to yield an observation probability map. The second stage generates the initial F0 trajectory using this map and signal power information. The final stage uses the deviation measure of each harmonic component and F0 adaptive time warping to refine the F0 estimate and aperiodicity estimation. The proposed framework is flexible to integrate other sources of instantaneous frequency when they provide relevant information.
Index Terms: fundamental frequency, speech analysis, speech synthesis, instantaneous frequency

1. Introduction

This paper describes a new F0 tracker for rapidly changing F0 trajectories with aperiodicity, which represents additive non-periodic components. In high-quality speech synthesis and modification applications [1–3], surpassing 4.2 on the 5 point MOS score, glitches in aperiodicity handling and the failure to follow rapidly changing fundamental frequencies (F0) are harmful to processed speech quality. Introducing a generative model of F0 trajectory (for example [4]) to F0 estimation provides well behaved and parametric representation. However, the estimated F0 trajectories are still not good enough for high-quality speech synthesis. The actual excitation signal of speech, glottal flow, contains several sources of fluctuations [5] and consequently, the observed F0 trajectories are different from the trajectories produced by those models. To attain highly natural synthetic speech it is important to retain these fine temporal variation in F0 trajectories [6, 7]. Although many F0 extractors have been proposed [8–12], in practice, parameter tuning and/or manual error correction is often necessary. In addition, their performance when extracting such fine temporal variations has not been investigated explicitly. That is the goal of this paper.

This paper is organized as follows. Section 2 discusses the motivation and target for designing a new F0 observer, based on a review on existing issues. It also defines aperiodicity, which is relevant for speech analysis and synthesis. Section 2.2 presents objective measures used in this paper. Based on these, section 3 introduces a general scalable architecture for F0 observer. It consists of three subsystems: front end aperiodicity detectors, the best trajectory finder, and F0 initial estimate and refinement subsystem with aperiodicity extractor. Sub-sections 3.1 and 3.3 introduce the front end and the refinement subsystems, respectively. In section 4, these subsystems are evaluated using artificial test signals. Section 5 discusses remaining issues. Example analysis results using actual speech samples and mathematical details are given in appendices.

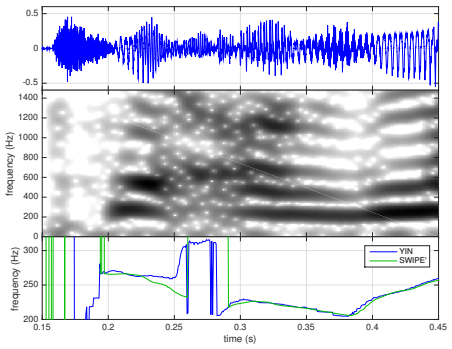

Figure 1: Example of the difficulty of handling irregular voicing. Upper plot shows speech waveform. Middle shows spectrogram using 25 ms Blackman window with 1 ms frame shift. Lower plot shows F0 trajectories extracted using YIN and SWIPE′. Around 0.25 s to 0.3 s, deviations caused discrepancies and/or failure of the baseline F0 trajectory trackers.

2. Background

Speech synthesis requires dependable F0 values whenever producing voiced sounds. However, even for copy-synthesizing from actual speech samples, where the targets are known, this is not always easy, since voiced sounds are not purely periodic and defining F0 values for such signals is not a trivial issue.

Figure 1 shows a beginning of a sentence from our speech corpus. From 0.2 s to 0.52 s, the speech signal is voiced. However, due to irregularities in glottal vibrations, defining the F0 is difficult. The lower plot shows the F0 tracks by YIN [10] and SWIPE′ [12] to illustrate the issues. It is difficult to evaluate the relevance of these tracking results. Yet these two state of the art systems do not produce consistent results. The fact that voicing without vocal fold contact is not rare [13, 14] prevents using EGG (electroglottograph) for the source of ground truth. Using the extracted trajectory and comparing the synthesized speech and the original speech is a reasonable test but it is very demanding on human resource and time to obtain reliable results.

An alternative approach for evaluating F0 extractors is to use an objectively defined artificial test signal. The ideal candidate is a speech signal, where the ground truth is available and provides wide divergence and variability. Instead, this article uses the excitation source signal defined by the L–F (Liljencrants–Fant) model [15]. The L–F model represents the time derivative of the glottal flow using a set of equations with four parameters. However, directly digitizing the L–F model, which is defined in the continuous time domain, introduces spurious components due to aliasing. To alleviate this aliasing problem this paper uses a closed-form representation of the anti-aliased L–F model defined in the continuous time domain [16].

http://dx.doi.org/10.21437/SSW.2016-36

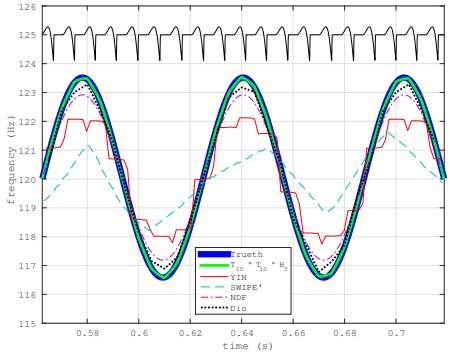

Figure 2: Frequency modulated F0 tracking. Black thin line on top shows waveform of the L–F (Liljencrants–Fant) model [15, 16] output. The very thick blue line shows the true F0 trajectory, which was used to generate the test signal. The refined F0 trajectory by the proposed method (thick light green line) almost overlays on the true trajectory.

Since the model is defined in the continuous time domain, it is easy to generate a signal using a given F0 trajectory that will be the ground truth used in this paper.[1] [20–22]

Figure 2 shows an example of F0 tracking using a sinusoidally frequency modulated F0 trajectory as the test signal. This test signal has a vibrato of 16 Hz, which is large compared to the normal human voice, but demonstrates the problems due to random, cycle-by-cycle variations in the F0. The tested F0 extractors are YIN [10], SWIPE′ [12], NDF [11], DIO [23] and the proposed method, which is described in Section 3. The trajectories obtained by YIN and SWIPE′ are strongly distorted and attenuated, perhaps because the F0 is changing faster than these models allow. When these distorted trajectories are used to generate the excitation source for copy-synthesis, the output is perceived differently. This is because the distortion adds fast-changing modulation components that are not in the original signal. The effects of these spurious components are made worse because humans are far more sensitive to fast frequency modulations than amplitude modulations [24, 25].

Voiced sounds are usually considered as periodic, and to first approximation the glottal pulses do occur at regular intervals. But due to prosodic needs the F0 of a voice is constantly changing, sometimes a simple glide as in the rise of F0 in a question, and sometimes in a regular fashion, as with vibrato. And, sometimes F0 varies in a more complex patterns, such as in tonal languages, where the F0 trajectory conveys linguistic information. On top of these intended changes in F0, there are modulations due to physiological aspects of voice production. The stochastic nature of neural pulses which drive the muscles of the vocal organ is a strong noise source and the critical conditions that produce vocal fold oscillation introduce bi-stable or chaotic vocal fold vibration, especially during voice onset and offset. Age related change and physical body status also affects the stability of vibration [5]. All these deviations from pure periodicity play important roles in speech communication and make speech a much richer media than text [26].

It is important to properly analyse and replicate these deviations from periodicity in high-quality speech synthesis and modification applications. Accurately estimating aperiodicity is still a very challenging problem. Tracking errors introduces spurious components [27, 28] and they add to the original random component. These are the reasons why F0 tracking distortions as shown in Fig. 2 are harmful for high-quality speech synthesis. Two issues have to be properly solved : accurate estimate and tracking of changing F0 trajectory and accurate estimate of random components based on the accurate estimate of F0 trajectory.

These issues motivate us to develop a framework that provides a calibrated procedure to describe the amount of aperiodicity and to track F0. The primary analysis target is high quality speech corpus recorded in a quiet and acoustically controlled environment using high-fidelity microphones. The aim here is to provide accurate, certified metadata, in this case, F0 value and an index that represents the accuracy of the estimated F0 as well as a measure that represents the amount of aperiodicity. Processing speed is not the first priority of the framework described here. Note that these metadata depend only on the data in the analysis frame, because there is no reliable model yet for the dynamic behaviour of F0 and aperiodic component. Using models of dynamic F0 behaviour such as Fujisaki's model [29], or F0 continuity constraint, may introduce biases due to model mismatch. Frame-based F0 with aperiodicity information, which the proposed system produces, will help to establish certifiably accurate models of the statistical/dynamic behaviour.

2.1. What is aperiodicity?

For speech synthesis applications, amplitude and F0 are controllable parameters of the excitation source. However, only replicating amplitude and F0 precisely to the original speech yields poor quality synthetic sounds. An important attribute of excitation is missing. This missing attribute is aperiodicity.[2]

In this paper, attributes that can be represented by amplitude and F0 modulation are not included in the definition of aperiodicity. What is left after removing periodic component defines "aperiodicity" in this paper. It turns out that our system's F0 estimation error is well correlated with the system's estimate of aperiodicity, described below.

2.2. Measures for objective evaluation

F0 extractors have been evaluated based on error-rate related measures; such as Gross Pitch Error (GPE), Voicing Detection Error (VDE) [30] and Pitch Tracking Error (PTE) [31]. Attaining high performance in these measures is a prerequisite for good F0 extractors. In this paper, we focus on F0 tracking fidelity, because the proposed method does not make voiced/unvoiced decision. Instead, this F0 tracker outputs a measure of aperiodicity, which closely correlates with the standard deviation of the relative F0 estimation error from the true value. This aperiodicity detector also is an informative source of the type of excitation. The voiced/unvoiced decision is left to the application, which can use the output of the proposed method to make this decision.

3. Architecture and subsystems

The proposed framework, YANGSAF (Yet ANother Glottal Source Analysis Framework), computes the instantaneous F0 using three steps: estimate, track, and refine. The estimation step calculates three features of the input signal over a number of bandpass channels. The maximum from the estimate stage is then tracked to produce a local estimate of the F0. Finally, an optional refinement stage combines temporal and harmonic information to produce a more accurate estimate of F0.

3.1. Estimation

The first stage of the YANGSAF algorithm analyses the signal with a number of bandpass channels, and then estimates three values for each channel as a function of time. These values are 1) the local instantaneous frequency, 2) a measure called aperiodicity that represents the amount of variability in the channel's frequency estimate, and 3) a probabilistic estimate that the channel contains a good representation of the F0. These signals are described in the subsections that follow and are used in the tracking stage described by Section 4.2. Figure 3 shows a diagram of the estimating detector in each channel.

The front end breaks the input into a number of spectral

[1] In an open-source implementation [17, 18] of the anti-aliased L–F model [16], the model parameters can be controlled each glottal cycle independently to simulate the details of vocal fold behaviour [19]. It can be combined with a time varying lattice filter to simulate the dynamic speech production process, which modulates observed F0 through interaction between harmonic component and the group delay associated with resonances (formant trajectories). But these detailed simulations are for further study.

[2] Effects of spectral envelope are also ignored. These details exceed the scope of this paper.

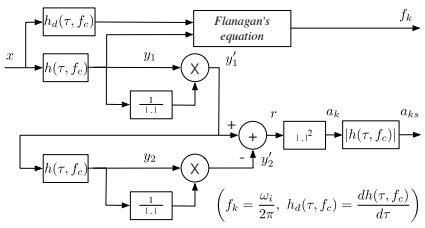

Figure 3: Schematic diagram of aperiodicity detector. Upper part calculates instantaneous frequency using Flanagan's equation (Appendix A). The lower part calculates aperiodicity measure as a relative residual level a_{ks} (Appendix B).

channels using a bank of bandpass filters, each centered at f_c.[3] The center frequencies cover the possible F0 range, with a fixed separation on the logarithmic frequency axis. The current implementation covers 400 Hz to 1000 Hz using 12 channels and detectors in each octave.

The instantaneous frequency estimate needs both the complex-valued signal and its derivative. These values are calculated starting with bandpass filter $h(\tau, f_c)$ and its derivative $h_d(\tau, f_c)$ shown in Fig. 3 and described in Appendix A. Each bandpass filter has linear phase, is a zero-delay FIR filter, has a complex-valued response, and passes only the positive frequency components.

Figure 4 shows an example of these three estimated signals for a sequence of vowels.

3.1.1. *Instantaneous Frequency*

The instantaneous frequency of the signal contained within each channel is calculated using Flanagsn's approach, which is based on the logarithm of a complex signal $x(t)$ and its derivative. An AM/FM modulated signal is represented in polar form $x(t) = r(t)e^{j\theta(t)}$. The instantaneous (angular) frequency $\omega_i(t)$ is defined as the derivative of the phase component $\theta(t)$, namely $\omega_i(t) = \frac{d\theta(t)}{dt}$. The instantaneous frequency can be derived by starting with the logarithm of the component phase and using a bit of algebra:

$$\frac{d\log(x(t))}{dt} = \frac{d\log\left(r(t)e^{j\theta(t)}\right)}{dt} = \frac{d\log(r(t))}{dt} + j\frac{d\theta(t)}{dt} \quad (1)$$

$$\omega_i(t) = \frac{\Re[x(t)]\frac{d\Im[x(t)]}{dt} - \Im[x(t)]\frac{d\Re[x(t)]}{dt}}{|x(t)|^2}, \quad (2)$$

where $\Re[x]$ and $\Im[x]$ represents the real and the imaginary part of x, respectively. The derivation of this expression is contained in Appendix A.

3.1.2. *Aperiodicity*

We also wish to calculate a measure of the aperiodicity of the signal in each channel, which will be used as a measure of the reliability of the instantaneous frequency measurement. For a constant sinusoid, the aperiodicity is zero, and the aperiodicity grows as the signal varies (wiggles) more within the bandpass channel. The basic idea of the periodicity detector is to calculate the amount of energy in the band-passed signal that is *not* the primary sinusoid. The primary sinusoidal component will have the largest energy, and when the complex signal is normalized to have unit magnitude, refiltered, and then renormalized, the primary sinusoid will still have unit magnitude. The other components will be filtered with a non-unit gain, since the filter is not an ideal brick-wall filter, and their amplitude will change. Subtracting the original and the twice-filtered and normalized response gives an estimate of the aperiodicity. Note this estimate is done *without* explicitly identifying the primary sinusoid and its frequency.

[3]The -3 dB points in frequency are $0.745f_c$ and $1.255f_c$. The zero points are located at 0 and $2f_c$. The -3 dB points in time are $-0.456/f_c$ and $0.456f_c$. Support is $(-2/f_c, 2/f_c)$.

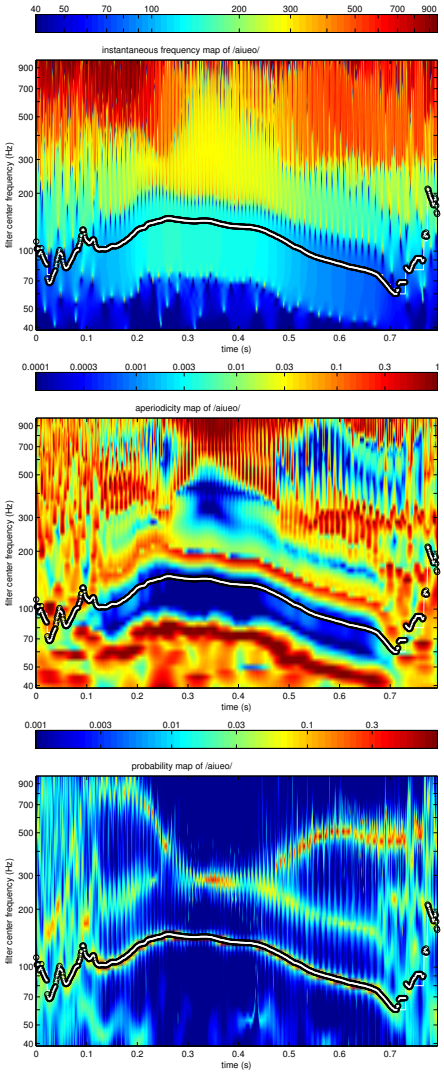

Figure 4: Example of the first stage detector outputs. The upper plot shows the instantaneous frequency map. The middle plot shows the residual map. The bottom plot shows the probability map. The speech material is a Japanese vowel sequence /aiueo/ spoken by a male. For reference purpose, the F0 trajectory extracted in the third stage is overlaid using open circles. In the probability map, the periodic vertical lines are synchronized with vocal fold vibration. The upper right trace of periodicity corresponds to the response of first formant of vowel /o/.

When a signal x whose fundamental frequency is equal to f_c is filtered, only the fundamental component, a complex-valued, slowly time-varying signal, is passed (appears in y_1) and is normalized to become y_1'. Then, by using the same filter, filtering signal y_1' again, and normalizing the overall amplitude using the absolute value of the complex valued-signal, the twice filtered (and amplitude normalized) signal y_2' is obtained. Subtracting this twice filtered and amplitude normalized signal y_2' from the amplitude normalized first filter output y_1' , yields a residual signal r. Since the signal y_1' is normalized, the power of the residual represents the relative level of the other component(s).

The difference between y_1' and y_2' corresponds to spectral components in the channel that are not the primary sinusoid. Calculating the energy in this signal (a_k), and smoothing it gives a_{ks} which is this system's measure of harmonic aperiodicity. Appendix B describes the relation between the SNR of the original signal and the residual aperiodicity power using

equations and examples.

Placing bandpass filters having the same shape on the logarithmic frequency axis yields the detector to output higher aperiodicity value, when f_c is located at harmonic frequencies other than the fundamental. This is similar to the concept "fundamentalness," which is explained in Fig. 11 of reference [32]. Appendix shows relation between filter shape examples and harmonic components.

The instantaneous frequency calculation and the aperiodicity calculation yield values at the audio sampling rate. These audio sampling rate time series are down-sampled for later processing. In this work the down-sampling is accomplished by extracting the nearest time samples from each time series, providing two sequences of instantaneous frequency and aperiodicity measure values at the frame rate (i.e. 200 Hz).

3.1.3. Probability

The fundamental component in the original signal is dominant in a number of output channels because there is little else for filters centered at frequencies lower than the second harmonic can respond. Thus a number of channels will respond in the same way to the fundamental component, as seen by the blueish blob around 100Hz in the second panel of Figure 4. All channels inside this blob have information about the fundamental component, but with different reliabilities.

Given a number of (distinct) estimates of the true F0, all from different channels, a probability map indicates which channel will have the best estimate. To create this probability map, all the instantaneous frequency and aperiodicity estimates are converted into Gaussian probability masses centered at various instantaneous frequency estimates. The output of the channel's aperiodicity estimate (a_{ks}, a measure of smoothed energy) is converted into a variance σ_k^2 by scaling. The scaling coefficient was empirically determined by a set of simulations. On a log-frequency scale ν, this gives a number of (independent) estimates of the instantaneous frequency, each modelled as a Gaussian mass centered at $\log(f_k)$, and with a variance of σ_k^2. Summing all these yields a probability density function $p_G(\nu)$ represented as a Gaussian mixture. For each channel, integrating this distribution provides an observation probability $P_r[k]$ that channel k should see the fundamental component in its nominal pass band $[f_L(k), f_H(k)]$ is

$$p_G(\nu) = \sum_{n=1}^{N} \frac{\hat{b}_n}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(\log(f_n)-\nu)^2}{\sigma_n^2}\right) \tag{3}$$

$$P_r[k] = \int_{\log(f_L[k])}^{\log(f_H[k])} p_G(\nu)d\nu \tag{4}$$

$$f_L[k] = f_c[k]2^{-\frac{1}{2K}}, \quad f_H[k] = f_c[k]2^{\frac{1}{2K}}, \tag{5}$$

where K represents the number of filters per octave.

This integrates the instantaneous frequency probability distributions between the frequency limits of filter k to arrive at an estimate of how reasonable it is for channel k to provide an estimate of the F0. An example of this result is shown in the bottom of Figure4.

3.2. Tracking

Given the three instantaneous maps (as a function of frame time and spectral channel) computed in Section 3.1, an initial estimate of the single best F0 at each frame is calculated by finding the channel with the highest probability. This is done in four steps: estimate the pitch range for this utterance, smooth the probability map, find the highest probability F0, and then refine the F0 estimate. The result is a smooth estimate of the true F0 based on the instantaneous frequency calculated in each channel.

First, the F0 search range is estimated by a weighted average of the instantaneous frequencies seen in the utterance. The temporal weighting is calculated from the energy in the original signal, after filtering it between 40-1000Hz, which is the prospective pitch range. Then each frame of the instantaneous frequency map is weighted and combined to form an overall instantaneous frequency histogram. By weighting by the signal's amplitude at each point in time, the high-energy portion of the utterance (vowels) are treated with more importance.

The median of this instantaneous frequency distribution (marginal distribution) defines the center point of the F0 search range. The tracker looks for peaks in the probability distribution within 1.2 octaves above this center point, and 1.3 octaves below, a total of a 2.5 octave range.

Second, in order to better estimate the F0 at the start and end of voicing the probability map computed in Section 3.1.3 is smoothed in time using a 45ms Hanning window with amplitude weighting. Smoothing is done before tracking so that we extend the F0 estimates at the start and end of voiced segments. For example, at the onset of voicing, the probability at F0 is not high, because the signal level is low and the SNR is low. Smoothing using amplitude weighting increases the probability at F0, because at frames after the onset the level grows and consequently the SNR become higher. In other words, the probability distribution of the onset frames become more like the probability distribution of later frames. This way smoothing reduces tracking error at the beginning of voicing. The same thing happens at the voice offset.

Thirdly, given the F0 range and the smoothed probability map the best channel across time can be tracked. For a range of channels that are within the 2.5 octave range defined for the entire utterance, and 0.7 octaves of the last frames best channel, the channel with the highest smoothed probability is chosen.

Finally, this channel selection is further refined by returning to the original probability map computed in Section 3.1.3 and choosing the channel with the highest probability closest to that bin chosen from the smoothed estimate. The following provides the initial F0 estimate f_{OI}.

$$f_{OI} = \sum_{m \in \mathbb{V}[k]} b_m f_m \tag{6}$$

$$\mathbb{V}[k] = \{ m \mid 0.5\, f_c[k] < f_c[m] < 1.25\, f_c[k] \}\ , \tag{7}$$

where the best weights b_m are calculated from σ_m^2 in $\mathbb{V}[k]$.

3.3. Refinement of the initial estimate

The third stage further improves this F0 estimate by adding two refinements. First, and most importantly, the higher harmonics of an F0 estimate can refine the estimate. Secondly, adaptive time warping of the original signal, combined with further refinement using higher harmonics of the warped signal, reduces the amount of F0 trajectory deviation for better analyses.

The first procedure uses harmonic frequencies and their variance. Each harmonic component, from first to m-th, has corresponding aperiodicity detector. Each bandpass filter of the detector has the same shape on the linear frequency axis and does not cover neighbouring harmonic components. Each detector yields instantaneous frequency f_k and its aperiodicity a_k, where k represents the harmonic number. These values are converted to F0 estimate f_k/k and its variance σ_k^2. The weighted average $\sum_{k=1}^{m} b_k f_k/k$ provides the refined F0 estimate. Variance values $\{\sigma_k^2\}_{k=1}^{m}$ are used to calculate the best mixing weights $\{b_k\}_{k=1}^{m}$ (Appendix D).

However, this refinement does not properly make use of higher harmonic information when the F0 trajectory is rapidly changing. This is because a rapid movement of higher frequencies generates strong side-band components and they smear the analysed harmonic structure [27, 28, 33].

Thus, the second procedure uses F0 adaptive time axis warping to alleviate this problem. Stretching the time axis, proportional to an instantaneous F0 value makes the observed F0 value constant [27, 28, 33] and keeps the harmonic structure intact. Then, placing aperiodicity detectors on harmonic frequencies, from first to m-th, the weighted average of F0 information yields the F0 estimate on the warped time axis. Converting this estimate value to the value on the original time axis provides the further improved F0 estimate.

These two procedures are applied serially as well as recursively. Let $\mathcal{H}_m$ represent the operation of harmonic based refinement using the first through m-th harmonic components and

$\mathcal{T}_m$ represent the operation of F0 adaptive time warping-based refinement using the first through m-th harmonic components. Let $\mathcal{P}_X[x;\Theta]$ represent the function of initial estimate F0 where x represents the input signal and Θ represents a set of the associated design parameters for analysis. The following equations describes the configurations of the two trackers tested:

$$\mathcal{H}_{10} \circ \mathcal{H}_3 \circ \mathcal{P}_X[x;\Theta] \tag{8}$$

$$\mathcal{T}_{10} \circ \mathcal{T}_{10} \circ \mathcal{H}_3 \circ \mathcal{P}_X[x;\Theta], \tag{9}$$

where $\mathcal{T} \circ \mathcal{H}$ represents the composite function of the functions $\mathcal{T}$ and $\mathcal{H}$.

Finally, by placing aperiodicity detectors on all harmonic frequencies in the warped time axis, estimated SNR around each harmonic component provides the excitation source information for speech synthesis. Because any F0 trajectories on this warped time axis are constant in time, aperiodicity values which detectors output are consistent with the aperiodicity definition of this paper.

4. Evaluation using test signals

This paper uses two measures of performance. Most importantly, the standard deviation of the relative error tells us the total distortion of the estimated F0 trajectory from the ground truth. The second performance measure is the frequency-modulation amplitude transfer function (FMTF), which expresses how well a F0 tracker follows fast F0 modulations. The test signal uses sinusoidal modulation on the logarithmic frequency axis, since F0 dynamics is better described on the logarithmic frequency axis [29]. Consequently, both FMTF and distortion evaluation measures use logarithmic frequency to calculate their value.

The proposed algorithms are implemented using MATLAB and tested using synthetic signals. Only representative results are described below. In the following tests, the test signals were generated using the aliasing-free L–F model [16].[4] The sampling frequency f_s was 22050 Hz and the "modal" voice quality parameters [34] for the L–F model were used in the following examples.

We test this new F0 tracker in two different ways: additive noise and FM modulation.

4.1. Additive noise

Firstly, the quality of the F0 estimate in the face of additive white noise was tested using the configuration given by Eq. 8 ($\mathcal{H}_{10} \circ \mathcal{H}_3$). The F0 extractor for the initial estimate (Section 3.2) ($\mathcal{P}_X[x;\Theta]$) was tested to clarify the effects of refinement (Section 3.3). Four popular F0 extractors were also evaluated for reference; YIN [10], SWIPE′ [12], NDF [11] and DIO [23, 35]. They were tested using their default or recommended settings. A constant F0 trajectory was used in this test.

Figure 5 shows the results for a 120 Hz F0. The vertical axis represents the relative RMS error. When the SNR is larger than 5 dB, YIN yielded the best results. But, YIN's performance is obtained at the cost of poor temporal resolution, which will be shown in the following test. DIO was designed for high-quality recordings and is not tolerant to noise. While SWIPE′ showed good performance from 0 to 20 dB SNR, performance saturated there after. The harmonic refinement procedure reduced the error in the initial estimate by a factor of 8, even in high noise, because the standard deviation of error in n-th harmonic component is $1/n$ as described in previous paragraph. In total, this is the second best result.

4.2. Frequency modulation of F0

Measuring the ability of a F0 tracker to follow F0 modulation is a more relevant test for speech sounds with rapid changes. The instantaneous frequency of the aliasing-free L–F model output was controlled at audio sampling rate (22050 Hz) resolution.

[4]The original L-F model [15] is anti-aliased using a closed form representation. The MATLAB implementation of this function and GUI-based interactive application for speech science education are open source [17, 18]. Spurious levels around the fundamental component of the model's output are lower than -120 dB.

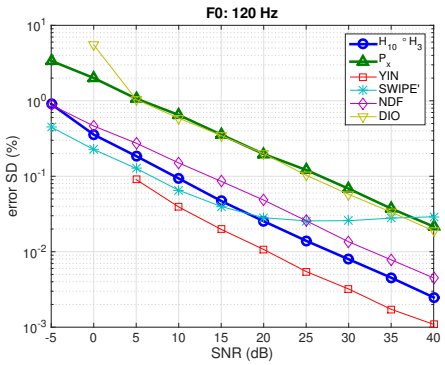

Figure 5: RMS error of F0 estimation vs. additive noise SNR for a temporally constant F0. The initial estimate (triangle) error deviations were reduced by a factor of 8 (circle) by using harmonic refinement.

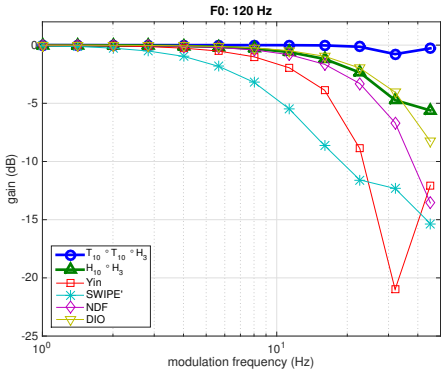

Figure 6: Frequency modulation transfer function for F0 modulation. The higher tracking frequency limit of the initial F0 estimate (triangle) is expanded two times by the proposed refinement using F0 adaptive time warping (circle).

The average F0 was 120 Hz with 100 musical cent peak-to-peak modulation depth roughly to 6% frequency modulation peak-to-peak in frequency.[5] In the two tests described in this section, a bit of white noise (SNR 100 dB) was added.

Figure 6 shows the frequency modulation transfer function for the four F0 trackers that serve as a benchmark and two variations of the F0 tracker described in this paper. For very low vibrato frequency (low modulation frequency) all F0 trackers work well at high SNR. At higher modulation frequencies all F0 trackers except for $T_{10} \circ T_{10} \circ H_3$ fail to follow the full modulation, which shows up as a reduced gain when considering the output vs input modulation deviation. For higher F0 signals, the 3 dB point increased proportionally to the F0 value, except YIN.

Figure 7 shows the RMS error of the F0 trajectories as a function of the modulation frequency. The dashed line and dash dot line show the RMS error of the best approximation to the true F0 using piece-wise linear function with segment lengths 1 ms and 5 ms respectively.

SWIPE′ and YIN yielded large RMS error, corresponding to the strong distortion shown in Fig. 2. The refinement performance without time warping is comparable to NDF. DIO showed the best performance among popular methods. The refined F0 trajectory using F0 adaptive time warping reduces the RMS error by a factor of 10 or more over the range from 2 Hz to 16 Hz modulation. For higher F0 values, RMS errors of other methods decrease inversely proportionally to the F0 value.

The F0 adaptive time warping also reduced spurious component due to FM substantially. For example, for a test signal with 16 Hz frequency modulation and 100 musical cent p-p depth, the refined F0 by the analysis configuration $\mathcal{T}_{10} \circ \mathcal{T}_{10} \circ \mathcal{H}_3$

[5]Tested F0 were 120, 240, 480 and 800 Hz. For F0 extractors, 120 Hz is the worst condition in terms of tracking.

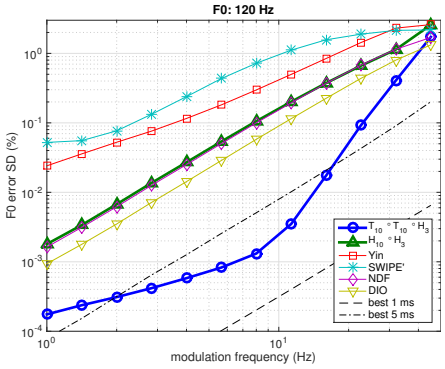

Figure 7: RMS error of F0 trajectory tracking. The RMS error of the refined F0 trajectory using harmonic frequencies (triangle) is reduced by a factor of 10 or more by introducing F0 adaptive time warping (circle).

reduced spurious residual levels lower than -40 dB. This is perceptually negligible.

5. Discussion

The goal of this paper is to estimate F0 trajectories, which consist of rapidly changing components, accurately for high-quality speech synthesis. The proposed set of procedures provide a prospective framework. However, the following aspects of F0 estimation were not exploited here. Investigations of the following issues could be important for improving synthesis quality further.

Plosive sounds such as /k/, /t/ sometimes sound like fricative by smearing temporal sharpness due to the smoothing effect of time windowing. This is a common degradation found in STRAIGHT.

Some speakers and languages frequently use "creaky voice." Representing these sounds using periodic signal plus noise results in poor reproduction. Relevant analysis and representations have to be investigated.

Temporal variation of F0 consists of effects caused by interactions between harmonic components and group delay in vocal tract transfer function. It is desirable to compensate this effect for speech synthesis applications, because this effect can be accumulated in each analysis and synthesis cycle.

In addition, it is interesting to consider a unique F0 tracker based on *Harmonic-Locked Loop* tracking [36] as an alternative F0 refinement procedure for the third stage of the proposed framework.

6. Conclusions

This paper introduced a framework for intantaneous estimates F0 and aperiodicity. It is able to improve the ability of F0 extractors to temporally follow varying F0 trajectories by a factor of 10. It may serve as an useful infrastructure for speech research and applications.

7. Acknowledgements

The authors appreciate insightful discussions with Prof. Roy Patterson on human auditory perception, especially on fine temporal structure and detection of interfering sounds. Malcolm Slaney provided editorial assistance. He and Dan Ellis also provided productive as well as critical comments.

A. Note on the Flanagan's equation

Flanagan uses the time derivative of the logarithm of a complex signal $x(t)$ to estimate the instantaneous frequency. By introducing a logarithmic function, the phase component is linearly separable from amplitude.

$$\log(x(t)) = \log(r(t)\exp(j\theta(t))) = \log(r(t)) + j\theta(t) \quad (10)$$

$$\Im[\log(x(t))] = \theta(t). \quad (11)$$

To make derivation simpler, as far as no ambiguity is introduced, time dependency representation by (t) is omitted afterwards.

$$\begin{aligned}
\omega_i = \frac{d\theta}{dt} &= \Im\left[\frac{d\log(x)}{dt}\right] = \Im\left[\frac{1}{x}\frac{dx}{dt}\right] \\
&= \Im\left[\frac{\frac{da}{dt} + j\frac{db}{dt}}{a+jb}\right] \quad \text{where } x = a + jb \\
&= \Im\left[\frac{\left(\frac{da}{dt} + j\frac{db}{dt}\right)(a-jb)}{(a+jb)(a-jb)}\right] \\
&= \Im\left[\frac{a\left(\frac{da}{dt} + j\frac{db}{dt}\right) - jb\left(\frac{da}{dt} + j\frac{db}{dt}\right)}{a^2+b^2}\right] \\
&= \Im\left[\frac{a\frac{da}{dt} + ja\frac{db}{dt} - jb\frac{da}{dt} - b\frac{db}{dt}}{a^2+b^2}\right] \\
&= \frac{a\frac{db}{dt} - b\frac{da}{dt}}{a^2+b^2} = \frac{\Re[x]\frac{d\Im[x]}{dt} - \Im[x]\frac{d\Re[x]}{dt}}{|x|^2}. \quad (12)
\end{aligned}$$

Which is the Flanagan's equation.

The complex-valued signal x in Eq. 12 is a filtered output of $h(t)$. It is a function of the center frequency f_c and time. Let explicitly represent x using $X(\omega_c, t)$ and its time derivative using $X_d(\omega_c, t)$. Then the following holds.

$$\begin{aligned}
X(t,\omega_c) &= \int_{-\infty}^{\infty} h(\lambda)x(t-\lambda)d\lambda \\
&= -\int_{-\infty}^{\infty} w(\tau - t)\exp\left(j\omega_c(\tau - t)\right)x(\tau)d\tau \quad (13) \\
X_d(t,\omega_c) &= \frac{dX(t,\omega_c)}{dt} \\
&= -\frac{d}{dt}\left(\int_{-\infty}^{\infty} w(\tau - t)\exp\left(j\omega_c(\tau - t)\right)x(\tau)d\tau\right) \\
&= -\int_{-\infty}^{\infty}\left(-\frac{d\,w(\tau - t)}{dt} - j\omega_c w(\tau - t)\right)\cdot \\
&\qquad \exp\left(j\omega_c(\tau - t)\right)x(\tau)d\tau \\
&= \int_{-\infty}^{\infty} h_d(\lambda)x(t-\lambda)d\lambda, \quad (14)
\end{aligned}$$

where

$$w_d(t) = \frac{dw(t)}{dt} + j\omega_c w(t) \quad (15)$$

$$h_d(t) = w_d(t)\exp(j\omega_c t). \quad (16)$$

Substituting these two time windows $w(t)$ and $w_d(t)$ into Eq. 12 removes time derivatives:

$$\omega_i(t,\omega_c) = \frac{\Re[X(t,\omega_c)]\Im[X_d(t,\omega_c)] - \Im[X(t,\omega_c)]\Re[X_d(t,\omega_c)]}{|X(t,\omega_c)|^2}, \quad (17)$$

Note that the TKEO (Teager Kaiser Energy Operator [37]) is not relevant for estimating rapidly changing F0 trajectories, since it uses an approximation, which requires slowly changing AM and FM. Using Flanagan's equation is relevant, since it does not rely on this approximation.

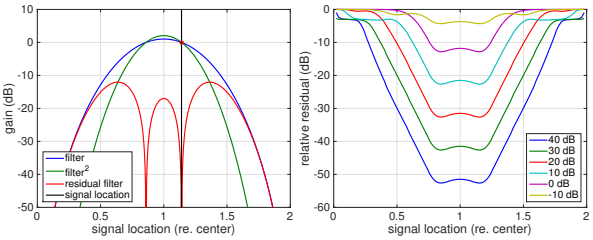

Figure 8: principles of operation. Left plot shows filter shape and the dominant signal at $1.14f_c$. Filter gains are adjusted to make output levels are 0 dB. Subtracting the second filter gain from the first one yields the equivalent filter for other components. Right plot shows the output residual level as a function of the location of the dominant signal and the noise level.

B. Residual calculation in each detector

This section shows how the aperiodicity detector in Fig. 3 works. The input to this detector is $x(t)$. Let $h(t, f_c)$ represent the complex valued impulse response of each band pass filter centered around f_c.

$$a_k(t, f_c) = |r(t, f_c)|^2 \tag{18}$$

$$r(t, f_c) = y_1'(t, f_c) - y_2'(t, f_c) \tag{19}$$

$$y_2'(t, f_c) = \frac{y_2(t, f_c)}{|y_2(t, f_c)|} \tag{20}$$

$$y_2(t, f_c) = \int_{-2/f_c}^{2/f_c} h(\tau, f_c) y_1'(t - \tau) d\tau \tag{21}$$

$$y_1'(t, f_c) = \frac{y_1(t, f_c)}{|y_1(t, f_c)|} \tag{22}$$

$$y_1(t, f_c) = \int_{-2/f_c}^{2/f_c} h(\tau, f_c) x(t - \tau) d\tau, \tag{23}$$

where the integration interval $(-2/f_c, 2/f_c)$ is for the Nuttall window (Eq. 25). For Hann window the interval is $(-1/f_c, 1/f_c)$ and for Blackman window the interval is $(-1.5/f_c, 1.5/f_c)$. Band pass filters having these impulse response lengths have first spectral zeros at 0 and $2f_c$.

Smoothing the relative residual level $a_k(t, f_c)$ yields the aperiodicity parameter $a_{ks}(t, f_c)$.

$$a_{ks}(t, f_c) = \int_{-2/f_c}^{2/f_c} |h(\tau, f_c)| a_k(t - \tau, f_c) d\tau. \tag{24}$$

B.1. Operation and implementation of the procedure

Figure 8 illustrates the process use to calculate the aperiodicity component. The impulse response of the filter $h(t, f_c)$ is

$$w(t) = \sum_{k=0}^{3} a_k \cos(2\pi k f_c t) \;\; |t| < \frac{2}{f_c} \tag{25}$$

$$h(t, f_c) = w(t) \exp(2\pi j f_c t), \tag{26}$$

where $j = \sqrt{-1}$ and the coefficients $\{a_k\}_{k=0}^3$ are (0.338946, 0.481973, 0.161054, 0.018027). This is the 11-th item in Table II of Nuttall's work [38].[6]

The detector is designed to cancel the primary periodic component in the input signal by adjusting the filter gain at the frequency of the primary component. This is done by normalizing the output by its RMS level. In a high SNR case, total RMS level of the filtered signals are approximately equal to the RMS level of the periodic component. The RMS level of the lower level components are affected by this suppression process. Since the equivalent filter gain from this suppression process is the difference of two filters, it yields the filter shape shown in the red curve of left plot of Fig. 8. The right plot of Fig. 8 shows the output aperiodicity parameter a_k as a function of the location of the primary component and the level of the lower level components.

[6] In terms of time-frequency product, when both is bounded, plorate spheroidal wave function is theoretically the best [39, 40]. However, due to large spectral dynamic range of actual speech signals, cosine series windows, which have very low side lobe level and steep side lobe decay [38] yielded better performance.

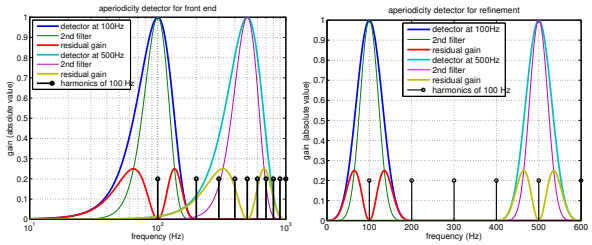

Figure 9: Detector allocation of the front end (left plot) and the refinement stage (right plot).

C. Detector allocation

Figure 9 shows detector filter shapes of front end and the third stage. In the front end, the filter width is proportional to the center frequency. In the refinement stage, the filter width is constant. The filters in the refinement stage are designed using the estimated F0.

D. Mixing F0 information

The band of estimators in the front end independently estimate instantaneous frequency and an estimate of the quality of this estimate in the form of an aperiodicity measure. We need to consolidate these estimates to get a single estimate of F0 and we do this with a weighted average.

Assume a set of random variables $X_k, k = 1, \ldots, N$ having zero mean ($\mathbb{E}[X_k] = 0$) and variances σ_k^2 ($\mathrm{Var}[X_k] = \sigma_k^2$). We wish to generate a new estimate from all the noisy estimates by weighting the individual estimates to arrive at an answer with the minimum estimated variance. Thus, assume the following cost function.

$$L = \mathrm{Var}\left[\sum_{k=1}^{N} b_k X_k\right], \tag{27}$$

where b_k represents the mixing coefficient. When mixing F0 estimates derived from different sources, the sum of weights has to satisfy the condition ($\sum_{k=1}^{N} b_k = 1$).

The optimum coefficients $\hat{b}_k$ for $k = 1, \ldots, N-1$ are derived by solving the following set of equations.

$$\sigma_N^2 = b_k \sigma_k^2 + \sigma_N^2 \sum_{n=1}^{N-1} b_n \quad \text{for } (k = 1, \ldots, N-1). \tag{28}$$

The final coefficient $\hat{b}_N$ is given by

$$\hat{b}_N = 1 - \sum_{k=1}^{N-1} \hat{b}_k. \tag{29}$$

Other source of F0 information can be used to improve this estimate further, if the variance of the estimate is available.

E. References

[1] Z. Heiga, T. Tomoki, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.

[2] H. Kawahara, M. Morise, Banno, and V. G. Skuk, "Temporally variable multi-aspect N-way morphing based on interference-free speech representations," in *ASPIPA ASC 2013*, 2013, p. 0S28.02.

[3] Y. Agiomyrgiannakis, "VOCAINE the vocoder and applications in speech synthesis," in *ICASSP 2015*, 2015, pp. 4230–4234.

[4] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1042–1053, 2015.

[5] I. R. Titze, *Principles of voice production*. National Center for Voice and Speech, 2000.

[6] T. Saitou, M. Unoki, and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech communication*, vol. 46, no. 3, pp. 405–417, 2005.

[7] L. Ardaillon, G. Degottex, and A. Roebel, "A multi-layer F0 model for singing voice synthesis using a B-spline representation with intuitive controls," in *Interspeech 2015*, 2015.

[8] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[9] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

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

[11] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT." in *Interspeech 2005*, 2005, pp. 537–540.

[12] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *JASA*, vol. 124, no. 3, pp. 1638–1652, 2008.

[13] D. Childers, D. Hicks, G. Moore, and Y. Alsaka, "A model for vocal fold vibratory motion, contact area, and the electroglottogram," *JASA*, vol. 80, no. 5, pp. 1309–1320, 1986.

[14] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *JASA*, vol. 87, no. 2, pp. 820–857, 1990.

[15] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech.*, vol. 4, pp. 1–13, 1985.

[16] H. Kawahara, K.-I. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and F0 extractor evaluation," in *APSIPA 2015*, Hong Kong, 2015.

[17] H. Kawahara. (2015) SparkNG: MATLAB realtime research tools for speech science education. [Online]. Available: http://www.sys.wakayama-u.ac.jp/%7ekawahara/MatlabRealtimeSpeechTools/

[18] ——, "SparkNG: Interactive MATLAB tools for introduction to speech production, perception and processing fundamentals and application of the aliasing-free L-F model component," in *Interspeech 2016*. ISCA, 2016, p. Show and Tell, (Accepted).

[19] K.-I. Sakakibara, H. Imagawa, H. Yokonishi, M. Kimura, and N. Tayama, "Physiological observations and synthesis of subharmonic voices," in *APSIPA ASC 2011*, 2011, pp. 1079–1085.

[20] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech.*, vol. 2-3, pp. 121–156, 1995.

[21] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *JASA*, vol. 90, no. 5, pp. 2394–2410, 1991.

[22] M. Garellek, G. Chen, B. R. Gerratt, A. Alwan, and J. Kreiman, "Perceptual differences among models of the voice source: Further evidence," *JASA*, vol. 136, no. 4, pp. 2295–2295, 2014.

[23] M. Morise, H. Kawahara, and T. Nishiura, "Rapid F0 estimation for high-SNR speech based on fundamental component extraction," *Trans. IEICEJ*, vol. J93-d, no. 2, pp. 109–117, 2010, [in Japanese].

[24] M. Tsuzaki and R. Patterson, "Jitter detection: A brief review and some new experiments," in *Proc. Symp. on Hearing, Grantham, UK*, vol. 53, 1997.

[25] C. C. Bergan and I. R. Titze, "Perception of pitch and roughness in vocal signals with subharmonics," *Journal of Voice*, vol. 15, no. 2, pp. 165–175, 2001.

[26] H. Fujisaki, "Prosody, models, and spontaneous speech," in *Computing Prosody*. Springer, 1997, pp. 27–42.

[27] T. Abe, T. Kobayashi, and S. Imai, "The IF spectrogram: a new spectral representation," *Proc. ASVA*, vol. 97, pp. 423–430, 1997.

[28] N. Malyska and T. F. Quatieri, "A time-warping framework for speech turbulence-noise component estimation during aperiodic phonation," in *ICASSP 2011*. IEEE, 2011, pp. 5404–5407.

[29] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, pp. 347–355, 1998.

[30] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *ICASSP 2009*. IEEE, 2009, pp. 3969–3972.

[31] B. S. Lee and D. P. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Interspeech 2012*. ISCA, 2012, pp. 707–710.

[32] 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.

[33] H. Kawahara, H. Katayose, A. D. Cheveigne, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," in *EuroSpeech'99*, 1999, pp. 2781–2784.

[34] D. G. Childers and C. Ahn, "Modeling the glottal volumevelocity waveform for three voice types," *JASA*, vol. 97, no. 1, pp. 505–519, 1995.

[35] 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 *AES 35*. Audio Engineering Society, 2009.

[36] A. L. Wang, "Instantaneous and frequency-warped techniques for source separation and signal parametrization," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 1995, pp. 47–50.

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

[38] 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.

[39] D. Slepian and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty-I," *Bell System Technical Journal*, vol. 40, no. 1, pp. 43–63, 1961.

[40] D. Slepian, "Prolate spheroidal wave functions, Fourier analysis, and uncertainty-V: The discrete case," *Bell System Technical Journal*, vol. 57, no. 5, pp. 1371–1430, 1978.