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

Improving YANGsaf F0 Estimator with Adaptive Kalman Filter

Kanru Hua

University of Illinois, U.S.A.
khua5@illinois.edu

Abstract

We present improvements to the refinement stage of YANGsaf[1] (Yet ANother Glottal source analysis framework), a recently published F0 estimation algorithm by Kawahara *et al*., for noisy/breathy speech signals. The baseline system, based on time-warping and weighted average of multi-band instantaneous frequency estimates, is still sensitive to additive noise when none of the harmonic provide reliable frequency estimate at low SNR. We alleviate this problem by calibrating the weighted averaging process based on statistics gathered from a Monte-Carlo simulation, and applying Kalman filtering to refined F0 trajectory with time-varying measurement and process distributions. The improved algorithm, adYANGsaf (adaptive Yet ANother Glottal source analysis framework), achieves significantly higher accuracy and smoother F0 trajectory on noisy speech while retaining its accuracy on clean speech, with little computational overhead introduced.

Index Terms: Fundamental Frequency, Monte-Carlo Simulation, Kalman Filter

1. Introduction

This paper presents a method for refining an initial fundamental frequency (F0) estimation on speech signals, with the goal of reducing fine error and improving temporal resolution of the estimated F0 trajectory. Such an algorithm is useful in speech synthesis, as many speech modeling techniques (e.g. STRAIGHT[2], aHM[3]) assume highly accurate F0 estimation as input.

In recent years there is a rising trend of formulating the F0 estimation problem in a Bayesian framework and the convenience of expressing noise in probabilistic notations in general leads to significant improvement in noise-robustness. A large class of these approaches assume a time-domain signal model with additive Gaussian noise. A likelihood function of observing the input signal, can then be defined and converted to a posterior distribution of F0. For example, Nielsen *et al*.[4] used a harmonic plus noise model; F0 and model order are jointly determined with maximum-entropy prior distribution. More recently, Hajimolahoseini *et al*.[5] used a simple stochastic time-delay model with period length controlled by another stochastic system to track the pitch period; inference is carried out using particle filtering. Despite the flexibility in model configuration, an issue commonly associated with the signal modeling approach is the validity of the assumed signal model: a simple model may not well-represent the speech generation mechanism, while a complicated model requires exotic inference techniques, or otherwise being computationally intractable.

Another category of probabilistic approaches to F0 estimation do not directly model the speech signal, but rather attempt to estimate F0 from features computed from the speech. Garner *et al*.[6] proposed a simple F0 tracker based on Kalman filter, where the measurement distribution is determined from the autocorrelation coefficients. Tsanas *et al*.[7] considered combining numerous F0 estimation algorithms also using a Kalman filter, but with measurement noise variance converted from a high level feature of algorithmic robustness of each estimator. However, the conversion from speech features to probability distribution is often based on empirically-designed heuristics which are not guaranteed to faithfully represent the statistics, and it involves lots of manually-specified parameters, which may overfit to the speech dataset used for parameter tuning.

It is also interesting to consider non-probabilistic and hybrid methods. Recently Stöter *et al*.[8] proposed a F0 refinement method based on iteratively time-warping the speech signal and updating F0 estimate on time-warped speech, which has a nearly constant F0. Taking it a step further, Kawahara *et al*.[1] embedded a feature-based probabilistic F0 estimator into the iterative time-warping paradigm, yielding the refinement stage of YANGsaf, the baseline method in this study. YANGsaf significantly improves the accuracy on clean speech but on breathy and noisy signals, especially when most of the harmonics are subjected to additive noise, a noisy fluctuation and occasionally spurious outliers, are observed in the refined F0 trajectory.

Goals of this research: we improve the noise-robustness of YANGsaf in the following direction. As opposed to the signal-modeling approaches, we focus on modeling the behavior of feature extractors under noisy conditions. Concretely, the relation between aperiodicity features and uncertainty of F0 estimate is studied based on synthetic signals generated from a Monte-Carlo simulation. The conversion from aperiodicity to frequency variance allows the adaptive smoothening of F0 trajectory using a Kalman filter. The outline of our approach is introduced in section 2. Section 3 presents the statistical analysis of feature extractors, toward the goal of estimating frequency variance from aperiodicity. In section 4 we apply adaptive Kalman filtering and subsequently Kalman smoothing on F0 trajectory. Section 5 evaluates the proposed adYANGsaf algorithm on speech signals at different noise levels. Finally our findings are summarized in section 6.

2. Outline of the proposed algorithm

The proposed algorithm (Figure 1) is an extension to YANGsaf, with improvements to the step converting aperiodicity to variance and a Kalman filtering step added.

The algorithm takes in a speech signal and an initial F0 estimation. The speech signal is first time-warped with respect to input F0 in such a way that attempts to remove the pitch fluctuation. The time-warped signal is processed by a filterbank that decomposes the signal by harmonics. For each harmonic, aperiodicity features (a ratio correlated with signal-to-noise ratio) and instantaneous frequency features are extracted at each frame. The method for feature extraction is described in [1]. The aperiodicity is then converted into a variance of instantaneous frequency, in terms of its relative deviation from the actual F0. Given the vector of frequencies and their variances, a weighted sum of harmonic frequencies can be easily com-

http://dx.doi.org/10.21437/Interspeech.2017-21

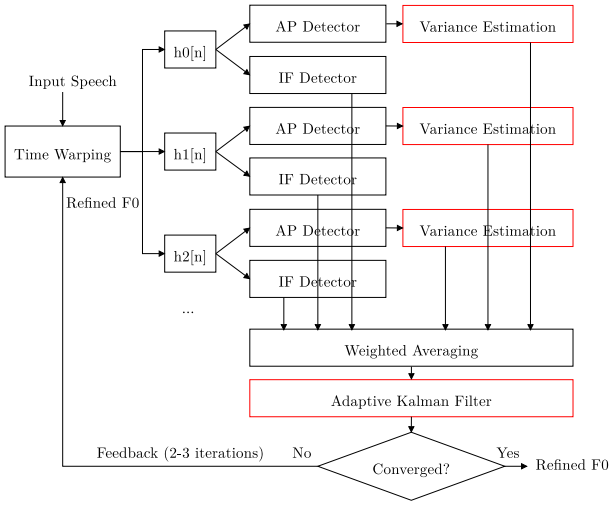

Figure 1: *Flowchart of the proposed algorithm. Additions and changes over the baseline are marked in red. "AP" and "IF" are the abbreviations for aperiodicity and instantaneous frequency, respectively.*

puted such that the variance of the sum is minimized, combining multiple frequency estimates into a most likely F0 estimate at each time instant. The new F0 trajectory is then smoothed by an adaptive Kalman filter, using the variances estimated in previous steps as parameters for the measurement distribution. According to Figure 1, the whole process is repeated for 2-3 iterations on the refined F0 trajectory.

It is worth noting that the original study[1] estimates the frequency variance by linearly scaling the aperiodicity feature. However, it remains a question regarding the relation between the aperiodicity extracted from bandpass filtered signal and the actual SNR. In addition, the extracted features can be distorted by noise, especially for higher harmonics, and the inaccuracy induced by such distortion has to be taken into account. In next section, the relation between aperiodicity features, the actual SNR and frequency variance is explored in detail.

3. Statistical analysis of feature extractors

3.1. Definitions

Seeing that the aperiodicity detector is not completely reliable, we need to distinguish the aperiodicity feature (referred to as raw aperiodicity $\mathbf{a_r}$ in the rest of this paper) from the actual SNR (referred to as true aperiodicity, $\mathbf{a_t}$) that is the ratio between noise standard deviation and sinusoid amplitude. The conditional distribution of observing raw aperiodicity given true aperiodicity is governed by the density function $p(\mathbf{a_r}|\mathbf{a_t}, f_r)$, where f_r is the constant for frequency resolution[1] of the bandpass filter. The other random variable conditioned on true aperiodicity is the instantaneous frequency, whose value randomly deviates from the harmonic due to noise. For simplicity a relative definition of instantaneous frequency deviation is used,

$$\mathbf{\Delta f} = \frac{\hat{\mathbf{f}} - f}{f_r} \tag{1}$$

where $\hat{\mathbf{f}}$ is the extracted instantaneous frequency and f is the assumed harmonic frequency. Our goal in the subsequent analysis is to find out the conditional distribution $\mathbf{\Delta f}|\mathbf{a_r}, f_r$, which provides a good estimation of variances of instantaneous frequencies.

[1]Frequencies are expressed as a ratio of sampling rate in this paper, unless otherwise noted

3.2. Monte-Carlo simulation for approximating the frequency variance

The small parameter space allows us to estimate the distribution of $\mathbf{\Delta f}|\mathbf{a_r}, f_r$ by running a Monte-Carlo simulation (Algorithm 1) over $\mathbf{a_t}$, repeated for different choices of f_r. Since all random variables ($\mathbf{\Delta f}$, $\mathbf{a_r}$ and $\mathbf{a_t}$) are scalars, it is relatively easy to approximate the result with an appropriate probability distribution, and to express the parameters of such distribution in terms of $\mathbf{a_r}$ and f_r, thus connecting $\mathbf{a_r}$ with the conditional variance $\sigma^2_{\Delta f|a_r,f_r}$. Two important prerequisites for the simulation are the prior distribution of raw aperiodicity $\mathbf{a_t}$ and the model for signal generation. In this study, the prior distribution is assumed to be log-uniform, between $[-50, 50]$ dB; we use a simple model adding a Gaussian noise with standard deviation $\sigma_g = \mathbf{a_t}$ to a sinusoid of amplitude 1 for drawing samples of the signal. Since the feature extractors operate on time-warped speech, we can safely assume the sinusoid has constant frequency, as the influence from slightly time-varying frequency is negligible compared to that from additive noise.

Algorithm 1: Monte-Carlo sample generation

Input: prior distribution of $\mathbf{a_t}$
list of frequency resolutions $f_r(k)$
number of samples N
Output: list of tuples (f_r, X), where X is a set of tuples $(a_r, \Delta f)$

```
for k = 1, 2, ..., K do
    set X = {}
    for n = 1, 2, ..., N do
        Sample a_t = a_t from its prior distribution
        Generate a test signal with σ_g = a_t
        Run aperiodicity and instantaneous frequency
          detectors on the test signal
        Add extracted features (a_r, Δf) to X
    end
    Store (f_r(k), X) for further analysis
end
```

For any particular choice of frequency resolution, once the samples are generated, a 2D histogram can be constructed from the list of $(a_r, \Delta f)$ tuples. Similarly the 1D histogram representing the conditional distribution $\mathbf{\Delta f}|\mathbf{a_r} = x, f_r$ can be estimated from a subset of the samples with $a_r \in [x - \epsilon, x + \epsilon]$, where ϵ is small compared to the range of $\mathbf{a_r}$. Figure 2 shows an example of such histogram along with a scaled normal PDF with the same variance as the empirical distribution. The empirical PDF is found to be bell-shaped but with greater kurtosis (sharpness) than a normal distribution.

The next step is to express $\sigma_{\Delta f|a_r,f_r}$, the conditional standard deviation of relative frequency deviation, as a function of $\mathbf{a_r}$ and f_r. We run the aforementioned simulation on 30 different f_r logarithmically spanned over the range $[0.0015, 0.045]$ with $N = 10^6$. Figure 3 shows the simulation result with all axes in logarithmic scale. Interestingly, the variance does not depend on frequency resolution. We find that $\log(\sigma_{\Delta f})$ is a linear function of $\log(\mathbf{a_r})$ for $\mathbf{a_r} \leq -70$ dB. For $\mathbf{a_r} > -70$ dB, the function has a sigmoidal shape and the standard deviation converges to $e^{-1.75} = 0.174$ at $\mathbf{a_r} = -30$ dB. The conver-

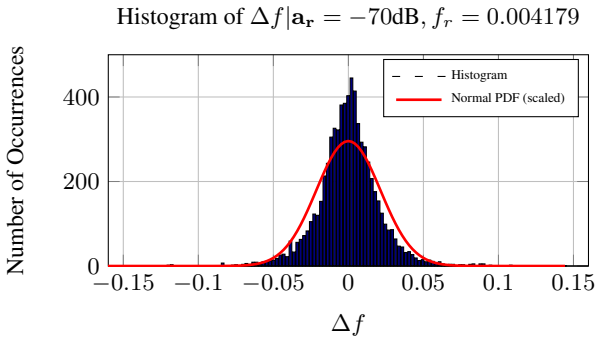

Figure 2: *Histogram of $\mathbf{\Delta f}|\mathbf{a_r} = -70dB$, $f_r = 0.004179$ overlaid with normal PDF with the same variance.*

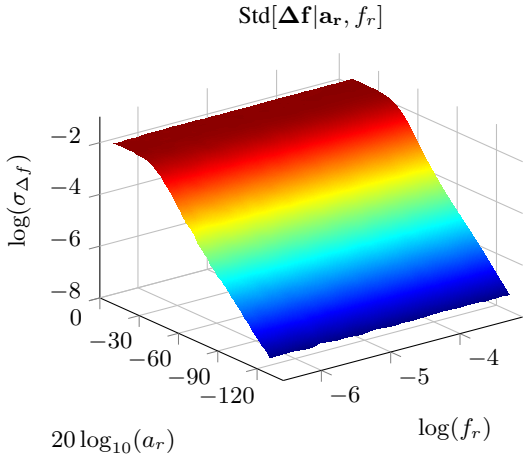

Figure 3: *surface plot expressing the standard deviation of $\mathbf{\Delta f}|\mathbf{a_r}, \mathbf{f_r}$ as a function of $\mathbf{a_r}$ and f_r. The maximum value of $\log(\sigma_{\Delta f})$ is -1.75.*

gence of Δf variance is consistent with the fact that the instantaneous frequency detector tends to pick up frequencies within the passband even if the input is completely random. We propose the following piecewise approximation to $\sigma_{\Delta f}$ as a function of $\mathbf{a_r}$,

$$\sigma_{\Delta f|a_r} = \begin{cases} \exp(0.206 + 0.0581 a_{r\text{dB}}), \\ \qquad\qquad a_{r\text{dB}} <= -70 \\ \exp(-1.75 - \frac{1}{0.382+\exp(6.687+0.129 a_{r\text{dB}})}), \\ \qquad\qquad a_{r\text{dB}} > -70 \end{cases} \tag{2}$$

Statistical analysis of the F0 refinement stage has led us to equation 2, a simple and useful result that can be directly used in place of the original frequency variance estimation method. The new variance estimator takes into account the uncertainties in the aperiodicity detector. When used in conjunction with a proper smoothing filter, the combined system should be more robust against additive noise.

4. Adaptive Kalman Filtering on Warped F0 Trajectory

The problem of recovering the optimal sequence of a hidden time series (the actual F0) from noisy observations (F0 estimate on each frame) naturally leads us to Bayesian filters, among which the most popular and efficient approach is Kalman filter. In the case of F0 tracking, we assume that the dynamics of F0 trajectory can be modeled as a single dimensional stochastic linear system,

$$x[t] = x[t-1] + w[t-1] \tag{3}$$
$$z[t] = x[t] + v[t] \tag{4}$$
$$w[t] \sim \mathcal{N}(0, Q[t]) \tag{5}$$
$$v[t] \sim \mathcal{N}(0, R[t]) \tag{6}$$

where $x[t]$ is the frequency at discrete time t; $w[t]$ is the random variable for process noise, which in our case represents the time-varying nature of F0. At each time instant a noisy measurement $z[t]$ is taken from the weighted average of several instantaneous frequency detectors. Inaccuracy of the weighted average is modeled as the additive measurement noise $v[t]$. The Gaussian assumption for measurement noise is justified by the fact that $x[t]$ is the weighted average of several noisy estimates with bell-shaped PDF, as shown in the previous section.

In general the model above is a scalar version of the state-space model used in Kalman filters, with the exception of both process and measurement noise distribution being time-varying. The measurement noise variance is the combination of variance estimates obtained in equation (2), scaled by the harmonic index, where b_k is the mixing weight of the k-th harmonic.

$$R[t] = \sum_k b_k[t] \frac{\sigma^2_{\Delta f|a_r,k}[t]}{k^2} \tag{7}$$

At this point the time-varying variance of process noise is unknown, and its identification is a non-trivial problem which will be explored in the next subsection.

When $Q[t]$, $R[t]$ and $z[t]$ are given, the Kalman filter estimates $\mathrm{E}[x[t]|z[1], ..., z[t]]$ at each time step[2]. To incorporate information about the whole observation sequence, a backward pass called Kalman smoothing is subsequently applied, giving an estimation of $\mathrm{E}[x[t]|z[1], ..., z[T]]$.

4.1. Identification of process noise variance

We consider the maximum-likelihood method [14] for identifying $Q[t]$, given $z[t]$ and $R[t]$. However the method does not generalize to cases with time-varying noise covariance. In addition, to prevent overfiting appropriate assumptions about $Q[t]$ have to be introduced. The key assumption is that $Q[t]$ is a scalar multiple of the moving variance taken from the input F0 trajectory,

$$Q[t] = \alpha \left(\frac{1}{2N} \sum_{k=-N}^{N} z^2[t+k] - \left(\frac{1}{2N} \sum_{k=-N}^{N} z[t+k] \right)^2 \right) \tag{8}$$

where the order of moving variance N is empirically found to be around 10. Due to inaccuracy of the feature extractors under noise, the measured trajectory $z[t]$ always has greater variance than the actual trajectory $x[t]$. Thus the scaling factor α is always in the range $[0, 1]$. Now that the search space is reduced to a bounded scalar value, a binary search on the log likelihood output from Kalman filter is possible.

[2]Since the posterior distribution is Gaussian, the expectation is also a maximum *a posteriori* estimation.

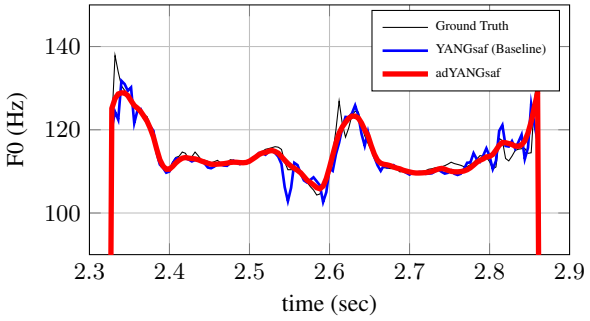

Figure 4: *Example of F0 estimated by adYANGsaf (red) and the baseline method (blue), plotted along with the ground truth.*

5. Evaluation

As all changes introduced to the baseline algorithm pertains to the refinement stage, our evaluation compares the fine error across several F0 estimation approaches, under different SNR levels.

5.1. Methodology and Settings

Twenty sentences (5 each from two male speakers and 10 from one female speaker) are selected from the CMU Arctic speech database [15]. The ground truth F0 is extracted from the EGG signal using YIN [16]. In addition, the squared difference estimated by YIN is thresholded at 0.01; frames with squared difference lower than 0.01 are marked as reliable and F0 fine error is collected only from reliable frames during evaluation.

Since the initial stage F0 estimator is prone to octave frequency error and voicing decision error at low SNR, it is desirable to eliminate its influence on the refinement stage. This is done by sending an artificially distorted version of ground truth to the refinement stage: a 10% Gaussian noise is added to ground truth F0 and a 30 ms moving average is taken twice on the result using the `filtfilt` Matlab function.

For testing noise robustness, Gaussian white noise and speech-weighted noise are added to the modified speech at several different SNR from 50 dB to -6 dB. The speech-weighted noise is generated by filtering the white noise signal by the average power spectrum of the speech estimated using the `pwelch` Matlab function, with a 128-tap Hanning window.

The following F0 estimation algorithms are tested along with adYANGsaf and the baseline method: RAPT[11], SWIPE[12], REAPER[13], YIN[16], DIO[9] and Harvest[10]. In all experiments F0 is extracted at 2 ms time step and fine error is calculated from all reliable frames on which the estimated F0 deviates from ground truth by less than 20%.

5.2. Results

Figure 5 plots the average F0 fine error of selected algorithms against different SNR for two types of noise. The error is displayed in dB, defined as $20\log_{10}\frac{\mathrm{std}(\mathbf{f_0}-\hat{\mathbf{f}}_0)}{\mathrm{std}(\mathbf{f_0})}$, where $\mathbf{f_0}$ and $\hat{\mathbf{f}}_0$ are the ground truth and estimated F0. The error of artificially dithered input F0 for adYANGsaf and the baseline method is shown as the dotted line. In both cases adYANGsaf consistently outperforms the baseline under all SNR, by a 5% to 40% margin and the improvement is more significant under low SNR conditions. The improved algorithm achieves lower fine error than a majority of tested algorithms in most conditions.

An example of F0 estimation on a male speech sample at 15 dB SNR is shown in Figure 4. The trajectory given by adYANGsaf is in general smoother and is not influenced by the dip at 2.55s, nor the fluctuation at 2.8s.

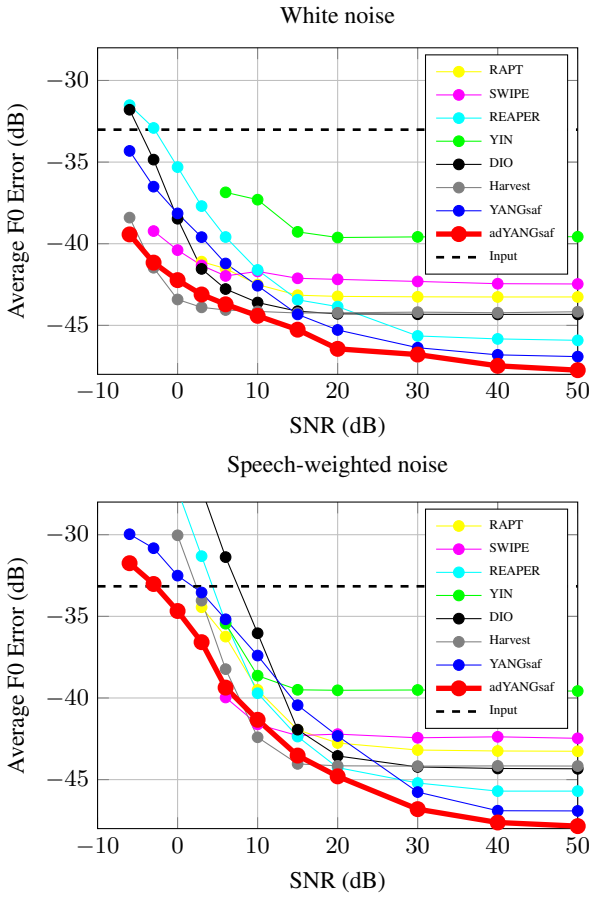

Figure 5: *Average F0 fine error of selected algorithms on speech corrupted by white noise and speech-weighted noise, respectively. The error of input F0 to adYANGsaf and the baseline method is shown as the dotted line.*

6. Conclusion

We proposed adYANGsaf, an improved version of YANGsaf by correcting aperiodicity-to-variance conversion and adding adaptive Kalman filtering to refinement stage. The improved algorithm consistently outperforms baseline in terms of F0 fine error, especially at low SNR scenarios. The probabilistic framework is effective for incorporating assumptions about the signal, and combining approaches from multiple disciplines. The result encourages further exploration on speech analysis algorithms based on modeling the characteristics of feature detectors through Monte-Carlo simulation.

7. References

[1] H. Kawahara, Y. Agiomyrgiannakis, and H. Zen, "Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis," in *9th ISCA Workshop on Speech Synthesis*, 2016.

[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] G. Degottex, and Y. Stylianou. "Analysis and synthesis of speech using an adaptive full-band harmonic model." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2085-2095, 2013.

[4] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "An Approximate Bayesian Fundamental Frequency Estimator," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE Press., 2012.

[5] H. Hajimolahoseini, R. Amirfattahi, S. Gazor, and H. Soltanian-Zadeh, "Robust estimation and tracking of pitch period using an efficient Bayesian filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1219–1229, Jul. 2016.

[6] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 102–105, Jan. 2013.

[7] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering," *The Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 2885–2901, May 2014.

[8] F. R. Stöter, N. Werner, S. Bayer, and B. Edler, "Refining Fundamental Frequency Estimates Using Time Warping," in *23rd European Signal Processing Conference (EUSIPCO)*, IEEE Press, 2015.

[9] 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].

[10] M. Morise, et al. "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Interspeech 2017*, [submitted].

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

[12] A. Camacho, "SWIPE: A sawtooth waveform inspired pitch estimator for speech and music," Phd diss, University of Florida, 2007.

[13] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," 2015. [Online]. Available: https://github.com/google/REAPER.

[14] R. L. Kashyap, "Maximum likelihood identification of stochastic linear systems," *IEEE Transactions on Automatic Control*, vol. 15, no. 1, pp. 25–34, Feb. 1970.

[15] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, Pittsburgh, 2004.

[16] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, p. 1917, 2002.