INTERSPEECH 2009 BRIGHTON

Model-Based Estimation of Instantaneous Pitch in Noisy Speech

Jung Ook Hong and Patrick J. Wolfe

Statistics and Information Sciences Laboratory
Harvard School of Engineering and Applied Sciences
33 Oxford Street, Cambridge, MA 02138

Abstract

In this paper we propose a model-based approach to instantaneous pitch estimation in noisy speech, by way of incorporating pitch smoothness assumptions into the well-known harmonic model. In this approach, the latent pitch contour is modeled using a basis of smooth polynomials, and is fit to waveform data by way of a harmonic model whose partials have time-varying amplitudes. The resultant nonlinear least squares estimation task is accomplished through the Gauss-Newton method with a novel initialization step that serves to greatly increase algorithm efficiency. We demonstrate the accuracy and robustness of our method through comparisons to state-of-the art pitch estimation algorithms using both simulated and real waveform data.

Index Terms: Harmonic model, instantaneous pitch estimation

10.21437/Interspeech.2009-26

1. Introduction

The pitch period is a function of the fundamental frequency of a voiced speech signal, and an important feature in speech analysis. A variety of pitch estimation algorithms exist in the literature, classified mainly according to their temporal assumptions on pitch variation. In the frame-based approach of [1], the pitch period is assumed constant within each analysis window; this forms the basis for pitch analysis algorithms in tools such as Praat [2] and WaveSurfer [3]. However, in reality, pitch cannot necessarily be assumed constant within each analysis window, leading to a number of problems. In regions of rapid pitch variation, for instance, shorter windows should be utilized, otherwise the resultant estimates are smeared. On the other hand, longer windows lead to more robust estimation in the presence of noise. In this paper, we propose a method which is able to track rapid pitch variation and is robust in low signal-to-noise environments.

Models of both the speech signal and the pitch process itself have been put forward in recent work, in order to increase the robustness and accuracy of pitch estimation. Tabrikian [4] introduced a "Harmonic plus Noise" model for modeling speech along with maximum a posteriori probability framework. The notion of instantaneous pitch, which assumes the process can vary continuously, has been used in the fine pitch model of Droppo [5] and the continuous pitch estimator of Resch [6]. Although [5] allows for discontinuous changes in pitch period from epoch to epoch, while [6] assumes further constraints, both share the same fundamental idea of the smoothness of the pitch process; i.e., that pitch periods tend to change in a relatively controlled manner over time.

2. The harmonic model of voiced speech

2.1. Harmonic model

We may model a segment of windowed voiced speech using the time-varying amplitude harmonic model [7]:

$$s(t) = \sum_{k=1}^{K} \{a_k(t)\sin(k\phi(t)) + b_k(t)\cos(k\phi(t))\}, \qquad (1)$$

where K is the number of harmonics, $a_k(t)$, $b_k(t)$ represent amplitudes of harmonic partials, and $\phi(t)$ is the instantaneous phase, i.e., $\phi(t) = \int_0^t \omega(\tau)d\tau$ where $\omega(t)$ is the instantaneous frequency of the windowed signal. With discrete time index $t \in \{t_1, t_2, \ldots, t_N\}$, we obtain discretized samples in the form of vector, i.e., $\boldsymbol{s} = [s(t_1), \ldots, s(t_N)]^T$ denotes a windowed speech segment of length N, with t_N on the order of 100 ms.

If the amplitude terms of (1) are assumed to be slowly time varying, we may represent them as $a_k(t) = \sum_{i=1}^{I} \alpha_{k,i}\psi_i(t)$ and $b_k(t) = \sum_{i=1}^{I} \beta_{k,i}\psi_i(t)$, where $\psi_i(t), i \in \{1, \ldots, I\}$ form a set of smooth basis functions; e.g., Hanning window function translated in time [7]. Denoting $\boldsymbol{\alpha} = [\alpha_{1,1}, \ldots, \alpha_{K,I}]^T$, $\boldsymbol{\beta} = [\beta_{1,1}, \ldots, \beta_{K,I}]^T$, and $\boldsymbol{\phi} = [\phi(t_1), \ldots, \phi(t_N)]^T$, $\boldsymbol{s}$ can be expressed in the form of a generalized linear model as follows:

$$\boldsymbol{s} = \boldsymbol{D}(\boldsymbol{\phi}) \left[\boldsymbol{\alpha}^T \boldsymbol{\beta}^T\right]^T, \qquad (2)$$

where elements of $\boldsymbol{D}(\boldsymbol{\phi})$ are constructed to satisfy the equation,

$$s(t) = \sum_{k=1}^{K}\sum_{i=1}^{I} \left[\psi_i(t)\sin(k\phi(t)) \;\; \psi_i(t)\cos(k\phi(t))\right] \begin{bmatrix}\alpha_{k,i}\\ \beta_{k,i}\end{bmatrix}$$

for $t \in \{t_1, \ldots, t_N\}$. We then adapt a standard additive noise model whereupon the vector of observations $\boldsymbol{y} = [y(t_1), \ldots, y(t_N)]^T$ is assumed to be corrupted by zero-mean white Gaussian noise with variance σ_e^2; that is $y(t_n) = s(t_n) + \epsilon(t_n)$ where $\epsilon(t_n) \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

2.2. Incorporating pitch dynamics

In frame-based pitch estimators, the instantaneous frequency term $\omega(t)$ is assumed to be $\omega_0 t$ within the analysis window of interest, where ω_0 is constant. In contrast, we adopt a model that exploits the notion of smoothness of the instantaneous pitch period; the pitch trajectory is modeled as a piecewise polynomial function of a low degree. This leads to an expansion of each piece of the pitch trajectory in a low-order polynomial basis.

To introduce this smooth pitch model, define L as the length of each piece, where it is assumed without loss of general-

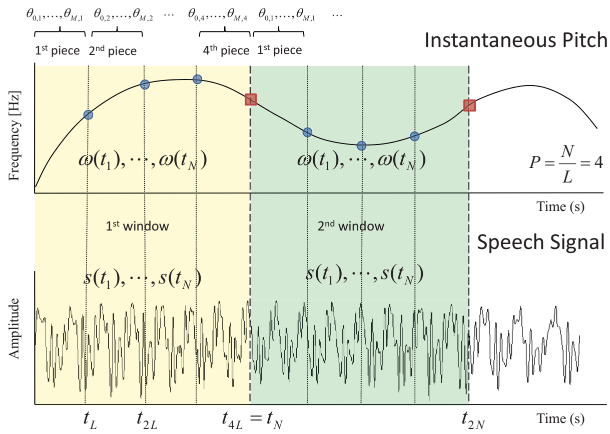

Figure 1: Diagram outlining the proposed model

ity that L divides N. Then, $P = \frac{N}{L}$ will be the number of pieces in one window. Let M denote the highest order of polynomial basis function employed, with the mth basis function given by $\eta_m(t)$, which is differentiable. $\eta_m(t)$ is then discretized at $t \in \{t_1, \ldots, t_L\}$ and we define the mth basis vector as $\boldsymbol{\eta}_m = [\eta_m(t_1), \ldots, \eta_m(t_L)]^T$. Also, let $\theta_{m,p}$ be the expansion coefficient corresponding to the mth polynomial basis function of the pth piece for $m \in \{0, \ldots, M\}$ and $p \in \{1, \ldots, P\}$. Then define $\boldsymbol{\theta}$ to be a collection of all $\theta_{\cdot,\cdot}$; i.e., $\boldsymbol{\theta} = [\theta_{0,1}, \ldots, \theta_{M,1}, \ldots, \theta_{0,P}, \ldots, \theta_{M,P}]^T$.

See Figure 1 for further explanation. The top figure shows the instantaneous pitch frequency while the bottom figure shows the corresponding speech signal. As shown in the figure, a windowed speech segment of length N has a corresponding instantaneous pitch trajectory of length N. This pitch trajectory consists of P pieces (here $P = 4$) of length L, each being smoothly connected to its neighbors according to the smoothness assumption to be discussed in the following section. Each piece is expanded in the polynomial basis of length L with expansion coefficients $\theta_{m,p}$.

Instantaneous pitch $\boldsymbol{\omega} = [\omega(t_1), \ldots, \omega(t_N)]^T$ is then

$$\boldsymbol{\omega} = \boldsymbol{\Omega}\boldsymbol{\theta}, \tag{3}$$

$$\boldsymbol{\Omega} = \begin{bmatrix} [\boldsymbol{\eta}_0 \ldots \boldsymbol{\eta}_M] & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \ddots & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & [\boldsymbol{\eta}_0 \ldots \boldsymbol{\eta}_M] \end{bmatrix}.$$

Here $\boldsymbol{\Omega}$ is a block-diagonal matrix of size $N \times (M+1)P$, and $\boldsymbol{\theta}$ is a vector of length $(M+1)P$ representing the corresponding coefficients of the polynomial basis. This leads us to write $\boldsymbol{s}$ in (2) in terms of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ as

$$\boldsymbol{s} = \boldsymbol{s}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{D}(\boldsymbol{\theta}) \left[\boldsymbol{\alpha}^T \boldsymbol{\beta}^T\right]^T, \tag{4}$$

where $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$ are the unknown parameters that describe a windowed speech segment, with K, M, I, and L known.

2.3. Enforcing smoothness of pitch

In order to exploit our assumption of pitch smoothness, we constrain the instantaneous frequency to be continuously differentiable at the boundaries between pieces as well as the boundaries between windows. These two types of continuity are explained in Figure 1 as well. The boundaries between pieces and windows are represented by blue circles and red squares, respectively. At a blue circle, the expansion coefficients of the two pieces adjacent to the boundary are constrained so that the trajectory of the two pieces can be connected smoothly. At a red square, the expansion coefficients of the first piece of the current window are constrained to equal the coefficients that expand the last piece of the previous window. In order to assure the continuity between two pieces, we constrain the instantaneous frequency to be continuously J-times differentiable at the boundaries of the pieces; e.g., $w^{(j)}(t_L) = w^{(j)}(t_{L+1})$ where $(\cdot)^{(j)}$ denotes the jth derivative with respect to t. The continuity constraints, with discretized $\eta_m^{(j)}(t)$ at time $t \in \{t_1, \ldots, t_L\}$ are:

$$\begin{aligned} \eta_0^{(j)}(t_L)\theta_{0,p} + \ldots + \eta_M^{(j)}(t_L)\theta_{M,p} \\ = \eta_0^{(j)}(t_1)\theta_{0,p+1} + \ldots + \eta_M^{(j)}(t_1)\theta_{M,p+1} \end{aligned}$$

for $j \in \{0, \ldots, J\}$ and $p \in \{1, \ldots, P-1\}$.

Similarly, the continuity between two windows is assured by constraining the instantaneous frequency to be continuously J-times differentiable at the boundaries, that is, $w_{pr}^{(j)}(t_N) = w_{cr}^{(j)}(t_1)$ where w_{pr} and w_{cr} denote the instantaneous frequency of the previous window and the current window, respectively. Let $\theta_{0,0}, \ldots, \theta_{M,0}$ be a set of known coefficients from the previous window. Similarly then, for $j \in \{0, \ldots, J\}$,

$$\begin{aligned} \eta_0^{(j)}(t_L)\theta_{0,0} + \ldots + \eta_M^{(j)}(t_L)\theta_{M,0} \\ = \eta_0^{(j)}(t_1)\theta_{0,1} + \ldots + \eta_M^{(j)}(t_1)\theta_{M,1}. \end{aligned}$$

We then collect the two types of continuity constraints in matrix form as a function of $\boldsymbol{\theta}$, as follows:

$$\boldsymbol{R}\boldsymbol{\theta} = \boldsymbol{c}, \tag{5}$$

where $\boldsymbol{R}$ and $\boldsymbol{c}$ consist of $\eta_m(\cdot), \eta_m'(\cdot), \ldots, \eta_m^{(J)}(\cdot)$ for $m \in \{0, \ldots, M\}$ at time t_1 and t_L, and $\theta_{0,0}, \ldots, \theta_{M,0}$ from the previous analysis window.

3. Parameter estimation

In our speech model (2) with the pitch model described in (3), the unknown parameters are $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$. We would like to estimate the unknown parameters by minimizing a cost function subject to (5). Writing the cost function $\mathcal{L}$ with Lagrange multipliers $\boldsymbol{\lambda}$ and the optimized parameters, $\boldsymbol{x} = \left[\boldsymbol{\alpha}^T \boldsymbol{\beta}^T \boldsymbol{\theta}^T\right]^T$, we have

$$\mathcal{L}(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{s}(\boldsymbol{x})||^2 + \boldsymbol{\lambda}^T(\boldsymbol{R}^*\boldsymbol{x} - \boldsymbol{c}^*) \tag{6}$$

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}), \tag{7}$$

where $\boldsymbol{R}^*$ and $\boldsymbol{c}^*$ are augmented $\boldsymbol{R}$ and $\boldsymbol{c}$ with zeros such that $\boldsymbol{R}^*\boldsymbol{x} = \boldsymbol{R}\boldsymbol{\theta}$ and $\boldsymbol{R}\boldsymbol{\theta} = \boldsymbol{c}$ if and only if $\boldsymbol{R}^*\boldsymbol{x} = \boldsymbol{c}^*$.

3.1. Gauss-Newton method

Realizing that (7) is a nonlinear least squares estimate with constraints, we employ the Gauss-Newton method in order to obtain the global minimum of (6) iteratively. Then the procedure of updating the parameters from $\hat{\boldsymbol{x}}_q$ to $\hat{\boldsymbol{x}}_{q+1}$, the estimates at iteration q and $q+1$, respectively, is expressed as

$$\hat{\boldsymbol{x}}_{q+1} = \arg\min_{\boldsymbol{x}} \{\frac{1}{2}||\boldsymbol{y} - \boldsymbol{s}_L(\boldsymbol{x}, \boldsymbol{x}_q)||^2 + \boldsymbol{\lambda}_q^T(\boldsymbol{R}^*\boldsymbol{x} - \boldsymbol{c}^*)\},$$

where $\boldsymbol{\lambda}_q$ is the Lagrange multiplier at iteration q, and $\boldsymbol{s}_L(\boldsymbol{x}, \boldsymbol{x}_q)$ is the linear approximation of (1) around $\boldsymbol{x}_q$:

$$\boldsymbol{s}_L(\boldsymbol{x}, \boldsymbol{x}_q) = \boldsymbol{s}(\boldsymbol{x}_q) + \boldsymbol{J}_q(\boldsymbol{x} - \boldsymbol{x}_q)$$

where $\boldsymbol{J}_q = \boldsymbol{J}(\boldsymbol{x}_q)$ is the Jacobian matrix of (2) with respect to $\boldsymbol{x}$. Note that $\hat{\boldsymbol{x}}_{q+1}$ and $\boldsymbol{\lambda}_{q+1}$ can be obtained by solving the Karush-Kuhn-Tucker matrix [8]:

$$\begin{bmatrix} \boldsymbol{J}_q^T \boldsymbol{J}_q & \boldsymbol{R}^{*T} \\ \boldsymbol{R}^* & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{x}}_{q+1} \\ \boldsymbol{\lambda}_{q+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_q^T(\boldsymbol{y} - \boldsymbol{s}(\boldsymbol{x}_q) - \boldsymbol{J}_q \boldsymbol{x}_q) \\ \boldsymbol{c}^* \end{bmatrix}.$$

3.2. Initial estimator

To initialize our iterative estimator, we require an initial estimator for the instantaneous frequency which should be sufficiently accurate to ensure that the iteration converges to the global minimum as expected, regardless of the presence of many local minima. Stoica [9] shows that the standard deviation of an initial frequency estimate should be on the order of $1/N$ in order that the Gauss-Newton method is able to find the global minimum, by showing the width of the valley of the global minimum is in the range $[2\pi/N, 4\pi/N]$, when N is the number of samples. Quinn [10] proposed a fast estimator of multi-sinusoidal components in time series and showed that the standard deviation of the estimator of each sinusoidal component is on the order of $\frac{1}{N^{3/2}}$, which satisfies the condition for the Gauss-Newton method to work.

In [10], the signal is modeled such that it has a known number of different frequency components, ω_k's, which are assumed to be constant within the analysis window. Instead, we take harmonic structure into account by letting $\omega_k = \omega_0 \cdot k$, where ω_0 is the constant fundamental frequency (i.e. the pitch period). When time-invariant amplitudes and pitch period are assumed in (1), it can be verified that

$$s(t) \prod_{k=1}^{K} (1 - 2D\cos(\omega_0 \cdot k) + D^2) = 0$$

where D denotes the unit delay operator. This is equivalent to saying that $(1 - 2D\cos(\omega_0 \cdot k) + D^2)$ annihilates each kth harmonic partial in (1). Applying the annihilating filter to the noisy data $y(t_n) = s(t_n) + \epsilon(t_n)$, we obtain an $\mathrm{ARMA}(2K, 2K)$ model

$$\sum_{k=1}^{2K} \varphi_k y(t_{n-k}) = \sum_{k=1}^{2K} \varphi_k \epsilon(t_{n-k}), \qquad (8)$$

where the parameters $\{\varphi_k\}$ are determined by the polynomial

$$\sum_{k=0}^{2K} \varphi_k z^k = \prod_{k=1}^{K} (1 - 2z\cos(\omega_0 \cdot k) + z^2).$$

Thus, with the special ARMA model of (8), whose AR and MA parts are identical, the problem of estimating the fundamental frequency reduces to that of estimating the φ_k's. Using each side of (8), we can adopt the iterative least squares approach of [10] in order to obtain an initial frequency estimate $\hat{\boldsymbol{\theta}}_0$.

Algorithm 1 summarizes our pitch estimation algorithm for voiced speech signals whose length is greater than N, in which case more than one analysis window is needed. The algorithm will return a set of final estimates, $\hat{\boldsymbol{x}}_f$'s, whose number corresponds to the number of analysis windows needed for the entire signal.

Algorithm 1 Pitch Estimation Per Windowed Segment

1. Initialize $\hat{\boldsymbol{\theta}}_0$ by the modified Quinn's algorithm introduced in Section 3.2. Then obtain $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\beta}}_0$ from (4) by substituting $\hat{\boldsymbol{\theta}}_0$, and solve for $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\beta}}_0$.
2. Set $q = 1$ and start the Gauss-Newton iteration described in Section 3.1 with the initial estimate $\hat{\boldsymbol{x}}_0 = [\hat{\boldsymbol{\alpha}}_0^T \hat{\boldsymbol{\beta}}_0^T \hat{\boldsymbol{\theta}}_0^T]^T$. Update $q \leftarrow q + 1$ and iterate until convergence. If converged, stop the iteration and set the final estimate, i.e., $\hat{\boldsymbol{x}}_f = \hat{\boldsymbol{x}}_q$.

4. Results

We now present experimental results demonstrating the robustness and accuracy of our method relative to standard algorithms such as RAPT [3] and the maximum a posteriori estimator (MAP) of [4]. The RAPT algorithm, used in popular pitch analysis tools such as WaveSurfer, decides pitch candidates and their cost by searching for local maxima in the autocorrelation function of the windowed speech signal, and then using a dynamic programming technique, along with voiced/unvoiced transition costs to account for halving and doubling issues when estimating pitch period. To ensure the most useful comparison, we removed this unvoiced/voiced selection feature.

The MAP approach of [4] assumes a harmonic model for voiced speech so that each windowed signal is expressed with a generalized linear model whose basis functions depend on the fundamental frequency and the number of harmonic partials. The pitch track itself is determined by the MAP estimate of pitch track parameters over several windows. The estimator is implemented using a recursive dynamic programming technique, that chooses a predefined frequency grid as its states, which affects the resolution of estimated pitch period.

Experiments were performed using both simulated and real waveform test data, with the former generated by a cascade of second-order digital resonators. The model parameters used were $(K, I, M, \frac{N}{f_s}) = (5, 8, 4, 200\text{ ms})$ or $(5, 16, 4, 400\text{ ms})$ depending on the sampling frequency, f_s. Figure 2 shows the estimated pitch period from RAPT, MAP, and our voiced speech model (VSM) with the Gauss-Newton method, plotted with the true instantaneous pitch period used to generate the simulated signal. The dotted line represents linear interpolations of pitch period from a frame-based estimator. MAP performs better at low SNR's than high SNR's as mentioned in [4], where RAPT fails to estimate desired pitch period by choosing subharmonics. Our scheme shows consistent performance in both the high and low SNR regimes.

The performance of each estimator in terms of mean-squared error is reported in Table 1. Gross error (GE), which is defined as occurring when the estimate differs from the true value by more than 50 Hz, is shown as well. Root mean-square error (RMSE) is calculated using only those samples which do not contribute to gross errors. The accuracy and robustness of our pitch estimation scheme can be observed by noting that corresponding RMSE values are similar regardless of noise level, and on average are lower than those of RAPT and MAP. Although MAP exhibits its robust performance in reducing GE, the algorithm does not improve its RMSE in that the accuracy of the estimator depends on the frequency resolution of the grid used to optimize its parameters through dynamic programming.

Results from experiments with rapid pitch variation are shown in Figure 3; RMSE is reported in Table 2. For the experiments, parameters $(K, I, M, \frac{N}{f_s}) = (5, 8, 5, 100\text{ ms})$ were used. It is clear that MAP performs better in terms of RMSE, whereas our scheme shows better performance in terms of GE.

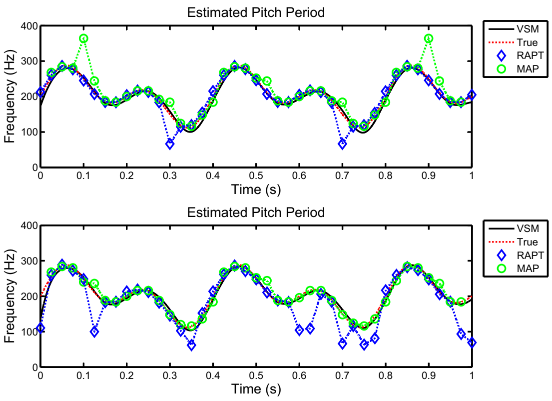

Figure 2: Slowly-varying pitch estimation results from simulated data, corrupted with noise to yield 15 dB (top) and 0 dB (bottom)

Table 1: Root mean-squared error and percentage of gross errors in estimated pitch, shown for simulated data with slow pitch variation

SNR (dB)	RMSE (Hz) (GE (%))		
	VSM	RAPT	MAP
15	8.74 (1.0)	9.36 (4.1)	14.19 (8.8)
10	7.65 (0.0)	7.93 (2.2)	12.87 (5.0)
5	6.64 (0.0)	8.78 (3.0)	12.08 (5.0)
0	8.77 (3.18)	14.02 (20.9)	11.73 (5.0)

Both our scheme and MAP succeed in estimating the rapid variation of pitch period, while RAPT fails in this case.

Figure 4 shows the estimated pitch period from a real waveform—a male-spoken utterance, "Why were you away a year Roy?". In the performance of MAP, the same tendency mentioned above is evident; it is the most robust, but least accurate estimator in terms of both GE and RMSE in high SNR situations. RAPT fails by choosing subharmonic and harmonic partials instead of the true pitch period. VSM computes estimates consistently across noise levels, failing only at the low-amplitude utterance end in low SNR.

To achieve analysis of a 1-second signal sampled at a rate of 16 kHz and degraded to 15 dB SNR, Matlab implementations of our proposed VSM estimator took 70 seconds while MAP and RAPT took 572 seconds and less than 1 second, respectively. For the 0 dB SNR case, VSM required 82 seconds.

5. Acknowledgements

Work supported in part by the Defense Advanced Research Projects Agency undera Grant HR0011-07-1-0007, and by the National Science Foundation under Grant DMS-0631636.

6. References

[1] P. Veprek and M. S. Scordilis, "Analysis, enhancement and evaluation of five pitch determination techniques," *Speech Comm.*, vol. 37, pp. 249–270, 2002.

[2] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci.*, 1993, pp. 97–110.

[3] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Paliwal, Eds. New York, NY: Elsevier, 1995, pp. 495–518.

[4] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 76–87, 2004.

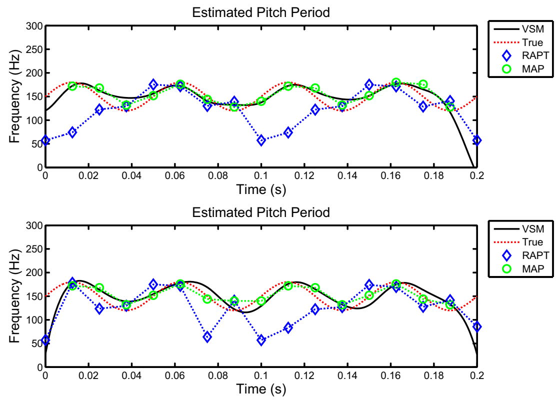

Figure 3: Rapidly-varying pitch estimation results from simulated data, corrupted with noise to yield 15 dB (top) and 0 dB (bottom)

Table 2: Root mean-squared error and percentage of gross errors in estimated pitch, shown for simulated data with rapid pitch variation

SNR (dB)	RMSE (Hz) (GE (%))		
	VSM	RAPT	MAP
15	15.75 (5.1)	18.81 (22.9)	12.67 (12.5)
10	15.62 (4.9)	19.10 (20.1)	12.76 (12.5)
5	15.46 (4.7)	19.52 (21.5)	13.20 (12.5)
0	15.16 (4.6)	22.25 (25.0)	15.08 (12.5)

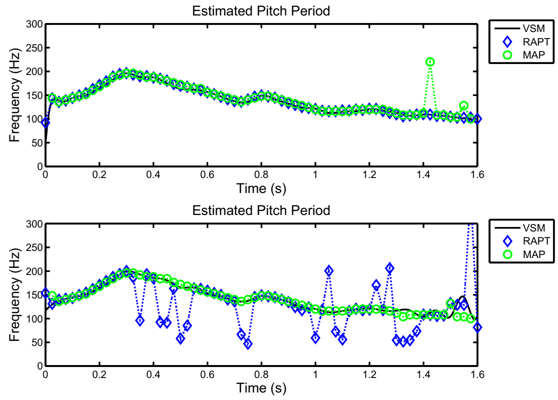

Figure 4: Pitch estimation results from real data corrupted with noise to yield 15 dB (top) and 0 dB (bottom).

[5] J. Droppo and A. Acero, "A fine pitch model for speech," in *Proc. INTERSPEECH*, 2007, pp. 2757–2760.

[6] B. Resch, M. Nilsson, A. Ekman, and W. B. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Trans. Audio Speech Lang.*, vol. 15, pp. 813–822, 2007.

[7] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, 2002, pp. 1769–1772.

[8] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY: Cambridge University Press, 2004.

[9] P. Stoica, R. L. Moses, B. Friedlander, and T. Soderstrom, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, pp. 378 – 392, 1989.

[10] B. G. Quinn, "A fast efficient technique for the estimation of frequency: interpretation and generalisation," *Biometrika*, vol. 86, pp. 213–220, 1999.