Improving the Accuracy and the Robustness of Harmonic Model for Pitch Estimation

Meysam Asgari and Izhak Shafran
Center for Spoken Language Understanding
Oregon Health & Science University
Portland, OR, USA
{asgari,shafrani}@ohsu.edu

Abstract
Accurate and robust estimation of pitch plays a central role in speech processing. Various methods in time, frequency and cepstral domain have been proposed for generating pitch candidates. Most algorithms excel when the background noise is minimal or for specific types of background noise. In this work, our aim is to improve the robustness and accuracy of pitch estimation across a wide variety of background noise conditions. For this we have chosen to adopt, the harmonic model of speech, a model that has gained considerable attention recently. We address two major weakness of this model. The problem of pitch halving and doubling, and the need to specify the number of harmonics. We exploit the energy of frequency in the neighborhood to alleviate halving and doubling. Using a model complexity term with a BIC criterion, we chose the optimal number of harmonics. We evaluated our proposed pitch estimation method with other state of the art techniques on Keele data set in terms of gross pitch error and fine pitch error. Through extensive experiments on several noisy conditions, we demonstrate that the proposed improvements provide substantial gains over other popular methods under different noise levels and environments.

Index Terms: fundamental frequency estimation, robust pitch estimation

1. Introduction
Pitch estimation algorithms typically consists of two stages. First, several candidate frequencies are estimated for each frame. Subsequently, from amongst the identified candidates, most probable pitch trajectory is estimated using Viterbi algorithm after applying smoothing constraints.

Various methods in time, frequency and cepstral domain have been proposed for generating pitch candidates. Since the literature on this topic is extensive, we limit our brief overview to popular or recent algorithms that are directly relevant to this work and the empirical evaluations reported in this paper. Praat obtains candidates from local peaks in either autocorrelation or normalized cross-correlation function [1]. YIN uses a autocorrelation-based squared difference function followed by post-processing techniques to calculate candidates [2]. Analogous to convolution in time domain, methods in frequency domain locate peaks in power spectrum. Hermes proposed an algorithm that estimates the \( f_0 \) by seeking the frequency that maximizes the summation of subharmonics on the power-spectrum [3]. This however ignores the information present in frequencies that are not harmonically related. To address this drawback, Sun proposed the Subharmonic to Harmonic Ratio algorithm (SHR), where the height of the peaks with respect to the valleys are considered [4]. Drugman cast this into regression framework where the residuals are also modeled [5]. Recently, Kawahara proposed a time-frequency method called TANDEM-STRaight for voice analysis and pitch extraction [6]. It first employs a power-spectrum estimation method called TANDEM that adaptively represents the signal and eliminates periodic temporal fluctuations. Then, pitch frequency is calculated using a fixed-point algorithm called STRAIGHT. Their time-frequency algorithm is computationally expensive.

Generally speaking, the above mentioned algorithms excel when the background noise is minimal or for specific types of background noise. However, pitch estimators are widely employed in diverse applications which can benefit from better accuracy and better robustness. For example, pitch tremors in early stages of Parkinson’s disease can be subtle and accurate estimator will be useful in automated screening tasks for the disease [7]. Likewise, robust estimators will be useful in extracting pitch features from noisy and unpredictable backgrounds to detect affect in widely deployed spoken-dialog systems [8] or for detecting subtle social cues in everyday conversations [9].

For our work, we chose to adopt, the harmonic model of speech, a model that has gained considerable attention recently. This model takes into account the harmonic nature of voiced speech and can be formulated to estimate pitch candidates with maximum likelihood criterion, as described briefly in Section 2. The straight forward application of this model however leads to certain types of systematic errors – halving and doubling errors. We propose a method in Section 3.1 to mitigate these error while choosing the candidates per frame. Typically, the number of harmonic components considered in the model are assumed to be constant and given. This is however not optimal and the choice needs to be guided by task specific conditions such as noise. We address this problem in Section 3.2 using the popular Bayesian information criterion. We empirically evaluate and characterize the proposed improvements to harmonic model using a series of experiments with several background noise types and at different signal to noise ratios. The results reported in Section 4 demonstrate that these improvements consistently outperform other competing algorithms. Finally, we summarize the contributions of this paper.

2. Speech Analysis Using Harmonic Model
2.1. Harmonic Model
The popular source-channel model of voiced speech considers glottal pulses as a source of period waveforms which is being
modified by the shape of the mouth assumed to be a linear channel. Thus, the resulting speech is rich in harmonics of the glottal pulse period. The harmonic model is a special case of a sinusoidal model where all the sinusoidal components are assumed to be harmonically related, that is, the frequencies of the sinusoids are multiples of the fundamental frequency [10]. The observed voiced signal is represented in terms of a time-varying harmonic component and a non-periodic component related to noise.

Let \( y = [y(t_1), y(t_2), \ldots, y(t_N)]^T \) denote the speech samples in a voiced frame, measured at times \( t_1, t_2, \ldots, t_N \). The samples can be represented with a harmonic model with an additive noise \( n = [n(t_1), n(t_2), \ldots, n(t_N)]^T \) as follow:

\[
\begin{align*}
s(t) &= a_0 + \sum_{h=1}^{H} a_h \cos(2\pi f_0 h t) + b_h \sin(2\pi f_0 h t) \\
y(t) &= s(t) + n(t)
\end{align*}
\]

where \( H \) denotes the number of harmonics and \( 2\pi f_0 \) stands for the fundamental angular frequency. The harmonic signal can be factored into coefficients of basis functions, the fundamental angular frequency. The harmonic components which are determined solely by the given angular frequency \( 2\pi f_0 \) and the choice of the basis function \( \psi(t) \).

\[
\begin{align*}
s(t) &= \left[ \begin{array}{c} 1 \\
A_r(t) \end{array} \right] \left[ \begin{array}{c} a_0 \\
\alpha \end{array} \right] \\
A_r(t) &= \left[ \begin{array}{c} \cos(2\pi f_0 t) \\
\cdots \\
\cos(2\pi f_0 Ht) \end{array} \right] \\
A_t(t) &= \left[ \begin{array}{c} \sin(2\pi f_0 t) \\
\cdots \\
\sin(2\pi f_0 Ht) \end{array} \right]
\end{align*}
\]

Stacking rows of \( [A_r(t), A_t(t)] \) at \( t = 1, \ldots, T \) into a matrix \( A \), equation (2) can compactly represented in matrix notation as:

\[
y = A m + n
\]

where \( y = A m \) corresponds to a expansion of the harmonic part of voiced frame in terms of windowed sinusoidal components, and \( \Theta = [f_0, b, \sigma_f^2, H] \) is the set of unknown parameters.

### 2.2. Pitch Estimation

Assuming the noise samples \( n \) are independent and identically distributed random variables with zero-mean Gaussian distribution, the likelihood function of the observed vector, \( y \), given the model parameters can be formulated as following equation. The parameters of vector \( m \) can then be estimated by maximum likelihood (ML) approach.

\[
L(\Theta) = -\frac{D}{2} \log(2\pi \sigma_n^2) - \frac{1}{2\sigma_n^2} ||y - Ab||^2
\]

\[
\hat{m}_{ML} = (A^T A)^{-1} A^T y
\]

Under the harmonic model, the reconstructed signal \( \hat{s} \) is given by \( \hat{s} = A \hat{m} \). The pitch can be estimated by maximizing the energy of the reconstructed signal over the pre-determined grid of discrete \( f_0 \) values ranging from \( f_{0_{\text{min}}} \) to \( f_{0_{\text{max}}} \).

\[
f_{0_{\text{ML}}} = \arg \max_{f_0} \hat{s}^T \hat{s}
\]

The pitch variations are inherently limited by the motion of the articulators in the mouth during speech production and hence they cannot vary arbitrarily between adjacent frames.

This smoothness constraint can be enforced using a first order Markov dependency between pitch estimates of successive frames. Adopting the popular hidden Markov model framework, the estimate of pitch over utterances can be formulated as follows. The observation probabilities are assumed to be independent given the hidden states or candidate pitch frequencies here. A zero-mean Gaussian distribution defined over the pitch difference between two successive frame is a reasonable approximation for the first order Markov transition probabilities [11].

\[
P(f_0^{(i)} | f_0^{(i-1)}) = \mathcal{N}(f_0^{(i)} - f_0^{(i-1)}, \sigma_f^2)
\]

Putting all this together and substituting the likelihood from the Equation 5, the pitch over an utterance can be estimated as follows.

\[
F_0 = \arg \max_{F_0} \sum_{i=0}^{M} s_i^T s_i f_0^{(i)} + \log \mathcal{N}(f_0^{(i)} - f_0^{(i-1)}, \sigma_f^2)
\]

Thus, the estimation of pitch over an utterance can be cast as an HMM decoding problem and can be efficiently solved using Viterbi algorithm.

### 3. Two Problems with Harmonic Models

#### 3.1. Pitch Halving and Doubling

Like in other pitch detection algorithms, pitch doubling and halving are the most common errors in harmonic models too. The harmonics of \( f_0/2 \) (halving) include all the harmonics of \( f_0 \). Similarly, the harmonics of \( 2f_0 \) (doubling) are also the harmonics of \( f_0 \). The true pitch \( f_0 \) may be confused with \( f_0/2 \) and \( 2f_0 \) depending on the number of harmonics considered and the noise.

In many conventional algorithms, the errors due to halving and doubling are minimized by heuristics such as limiting the range of allowable \( f_0 \) over a segment or an utterance. This requires prior knowledge about the gender and age of the speakers. Alternatives include median filtering and constraints in Viterbi search [12], which remain unsatisfactory.

We propose a method to capture the probability mass in the neighborhood of the candidate pitch frequency. The likelihood of the observed frames falls more rapidly near candidates at halving \( f_0/2 \) and doubling \( 2f_0 \) than at the true pitch frequency \( f_0 \). This probability mass in the neighborhood can be captured by convolving the likelihood function with an appropriate window. Figure 1 illustrates the problem of halving and demonstrates our solution for it. The dotted line shows the energy of the reconstructed signal \( \hat{s} \) for a frame. A maximum of this function will erroneously pick the candidate \( f_0/2 \) as the most likely pitch candidate for this frame. However, notice that the function has a broader peak at \( f_0 \) than at \( f_0/2 \). The solid line shows the result of convolving the energy of the reconstructed signal with a hamming window. In our experiments, we employed a hamming window with the length of \( f_{0_{\text{min}}} / 2 \) to \( f_{0_{\text{max}}} \) where \( f_{0_{\text{min}}} \) is the minimum pitch frequency. The locally smoothed likelihood shows a relatively high peak at the true pitch frequency \( f_0 \) compared to \( f_0/2 \), thus overcoming the problem of halving.

#### 3.2. Model Selection

Another problem with the harmonic model is the need to specify the number of harmonics considered. This is typically not known \textit{a priori} and the optimal value can differ in different noise conditions. Davy proposed a sampling-based method for estimating the number of harmonics [13]. Their approach is based on Monte Carlo sampling and requires computationally expensive numerical approximations. Mahadevan employs
the Akaike information criteria (AIC) for tackling the problem of model order selection [14]. Here, we follow a Bayesian approach trying to maximize the likelihood function given by:

$$\hat{H} = \arg \max_H p(y, \Theta_H)$$  \hspace{1cm} (7)

where $\Theta_H$ denotes the model constructed by $H$ harmonics. The likelihood function increases as a function of increasing model order and often leads to the overfitting problem. We adopt the Bayesian information criterion (BIC) as a model selection criterion, where the increase in the likelihood is penalized by a term that depends on the model complexity or the number of model parameters. For the harmonic model, we include a term that depends on the number of data points in analysis window $N$.

$$BIC(H) = -2 \log p(y, \Theta_H) + H \log N$$  \hspace{1cm} (8)

Thus, in our proposed model selection scheme, we compute the average frame-level BIC for different model orders, ranging from $H = 2, \ldots, H_{\text{max}}$. For a given task or noise condition, we choose the number of harmonic that minimize the average frame-level BIC. The Figure 2 shows the comparison between the average frame-level score computed by ML and BIC metrics as a function of number of harmonics. In this case, seven harmonics appear to be an optimal trade-off between increasing the likelihood and maintaining a parsimonious model.

4. Experiments

In our previous work [15] we addressed the problem of voiced/unvoiced detection and here we focus on pitch estimation problem. To ignore the effect of voicing decision errors on pitch estimation results, we assumed the voiced boundaries are given. We evaluate the performance of our proposed method on a task of estimating pitch frequency on the Keele dataset [16]. The data set contains 10 phonetically balanced audio files from 10 speakers, 5 males and 5 females. This dataset provides a reference pitch and voicing labels obtained from the simultaneously recorded laryngograph signal. For evaluation, we exclude frames for which the voicing label in the corpus is uncertain. The speech was recorded in noise free conditions and for testing the robustness of our algorithm we contaminated them with several types of additive noise at different SNRs using Filtering and Noise-adding Tool (FaNT) [17]. We configured FaNT with telephone speech characteristics using G.712 filter, a narrow-band telephone speech bandpass filter with a flat frequency response between approximately 300 to 3400 Hz. This filtering makes the task of pitch estimation more challenging compare to the full-band scenario due to the spectral attenuation at harmonics below 300 Hz.

The test were performed on the clean and noisy data at different noise level ranging from 0 db to 20 db in several noise environments including, restaurant, subway, white, ear, street, exhibition, babble, and airport taken from Aurora noise dataset [17].

We assessed performance of the methods using following measures [5]: gross pitch error (GPE), which is defined as the percentage of $f_0$ estimates that deviate more than 20% of the ground truth; and the fine pitch error (FPE) that is the mean absolute error computed for estimates that are bellow than 20% of reference $f_0$.

4.1. Experimental Evaluation

We compared the performance of our proposed method with the following pitch estimation methods – (a) STRAIGHT-TANDEM, based on the fixed-point analysis on modified power-spectrum [6]; (b) YIN, a template matching method with the autocorrelation function in time-frequency domain and ad hoc post-processing [2]; and (c) SHR, a method based on Sub-harmonics to Harmonic Ratio [4]. In all cases, the search for optimal pitch frequency was performed over a range from 80 Hz to 400 Hz was and the frame rate was fixed to 100/sec. The performance of the methods are compared using only the frames corresponding to the reference voiced frames.

4.2. Results

In noisy conditions, Table (1) reports the average errors over all SNR bins ranging from 0 db to 20 db. On both clean and noisy speech, HM-SL clearly outperforms all other approaches in terms of GPE except in white noisy condition where the smoothing appears unnecessary and the HM outperforms all.

In Table (2) HM approach outperforms others except for
Table 1: Comparison of the proposed method with other popular methods in terms of gross pitch error (GPE) under clean and different noisy condition. In noisy conditions, the table reports the average over all SNRs ranging from 0 db to 20 db.

<table>
<thead>
<tr>
<th>Method</th>
<th>clean</th>
<th>restaurant</th>
<th>subway</th>
<th>white</th>
<th>car</th>
<th>street</th>
<th>exhibition</th>
<th>babble</th>
<th>airport</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHR</td>
<td>4.18</td>
<td>15.61</td>
<td>13.79</td>
<td>12.84</td>
<td>15.61</td>
<td>16.66</td>
<td>14.03</td>
<td>16.11</td>
<td>16.65</td>
</tr>
<tr>
<td>S-T</td>
<td>2.05</td>
<td>12.03</td>
<td>9.72</td>
<td>9.17</td>
<td>21.58</td>
<td>20.52</td>
<td>7.40</td>
<td>13.37</td>
<td>19.37</td>
</tr>
<tr>
<td>HM</td>
<td>4.81</td>
<td>19.08</td>
<td>15.15</td>
<td>3.45</td>
<td>15.83</td>
<td>20.36</td>
<td>10.17</td>
<td>19.31</td>
<td>18.32</td>
</tr>
<tr>
<td>HM-SL</td>
<td>2.04</td>
<td>8.20</td>
<td>5.49</td>
<td>4.76</td>
<td>6.82</td>
<td>9.77</td>
<td>4.52</td>
<td>8.23</td>
<td>8.16</td>
</tr>
</tbody>
</table>

Table 2: Comparison of the proposed method with other popular methods in terms of fine pitch error (FPE) under clean and different noisy condition. In noisy conditions, the table reports the average over all SNRs ranging from 0 db to 20 db.

<table>
<thead>
<tr>
<th>Method</th>
<th>clean</th>
<th>restaurant</th>
<th>subway</th>
<th>white</th>
<th>car</th>
<th>street</th>
<th>exhibition</th>
<th>babble</th>
<th>airport</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHR</td>
<td>2.52</td>
<td>4.02</td>
<td>3.60</td>
<td>3.98</td>
<td>3.56</td>
<td>3.87</td>
<td>3.96</td>
<td>4.06</td>
<td></td>
</tr>
<tr>
<td>S-T</td>
<td>2.72</td>
<td>5.61</td>
<td>5.18</td>
<td>4.49</td>
<td>5.61</td>
<td>5.81</td>
<td>4.87</td>
<td>5.86</td>
<td>5.75</td>
</tr>
<tr>
<td>HM</td>
<td>2.22</td>
<td>3.63</td>
<td>3.00</td>
<td>2.80</td>
<td>3.31</td>
<td>3.44</td>
<td>3.09</td>
<td>3.59</td>
<td>3.58</td>
</tr>
<tr>
<td>HM-SL</td>
<td>2.62</td>
<td>3.87</td>
<td>3.27</td>
<td>3.04</td>
<td>3.46</td>
<td>3.98</td>
<td>3.30</td>
<td>3.86</td>
<td>3.78</td>
</tr>
</tbody>
</table>

SHR in restaurant noisy condition. As it is clear in table 2, HM-SL has a comparable performance to the HM. This may be explained by the fact that smoothing of the likelihood score may reduce the precision of the harmonics.

Figure 3: Gross pitch error (%) (top) and fine pitch error (%) (bottom) for all methods and averaged over all 8 noisy conditions.

The Figure 3 summarizes the overall gross pitch error and fine pitch error across all the noise conditions. The proposed model (HM-SL) outperforms the other popular methods substantially in gross pitch error consistently under all levels of noise conditions. The performance of HM-SL is also better than HM, which shows that the smoothing contributes to the performance gains. The model also performs better in fine pitch error when the noise level is high. At low noise levels, the proposed model degrades fine pitch estimate, which is not entirely surprising and is due to the smoothing. In fact, at low SNRs standard HM is sufficient and smoothing is not necessary.

5. Conclusions

In this paper, we have addressed two outstanding problems related to harmonic models in the context of pitch estimation. Like other pitch estimation algorithms, the harmonic model suffers from pitch halving and doubling. We propose a local smoothing function that exploits the fact that there is more energy in the harmonics near the true pitch than at the corresponding neighborhoods of half or double the pitch. We utilize a local smoothing function to include this energy and improve the robustness of the pitch candidates in each frame. The harmonic model requires specification of the number of harmonics. The optimal choice depends on the noise conditions. We adopt a BIC criterion and define a model complexity that allows us to estimate the number of harmonics for each noise condition. We estimate the optimal number of harmonics using the average BIC per frame. Together these improvements provide substantial gain over other popular methods under different noise types and levels.

6. Acknowledgements

This research was supported in part by NIH Award 1K25AG033723, NSF Awards 1027834, 0964102 and 0905095 and Google Faculty Award. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NIH or NSF.
7. References


