A Hybrid Fundamental Frequency Estimator for Singing Voice

By Hui-Ling Lu, vickylu@leland

Abstract

Fundamental frequency estimators are important in studying a singer's performance. In this report, a hybrid fundamental frequency estimator is developed to detect the variation of the fundamental frequency of singing voices. The hybrid estimator combines the maximum likelihood harmonic matching and the auto-correlation methods. Using the harmonic matching to refine the search range for the auto-correlation method, we can reduce the computation and achieve accurate estimates in fine resolutions. This system has been tested on clean recordings of singing voices successfully. Although only singing voices have been tested, the system should work on any harmonic sounds.

1. Introduction

The fundamental frequency of voice is defined as the rate of glottal pulses generated by the vibration of the vocal folds during voicing of segments. Several different fundamental frequency estimation methods have been reported for speech signals. Based on the quasi-periodic assumption, these methods can be summarized into three catalogs. One group of methods detects the timing of some time-domain features. Other time-domain methods use auto-correlation functions or difference norms to detect similarity between the waveform and a time-lagged version of itself. Another family of methods operates in the frequency domain, locating sinusoidal peaks in the frequency transform of the input signal. Time-domain methods using auto-correlation functions or difference norms are most popular and robust. A method belonging to this group, called SRPD (super resolution pitch determination method) [5], is widely used for the speech signals. We will use this method to refine the fundamental frequency estimates after a rough search range is obtained via the harmonic matching method that will be introduced in the next section.

The objective of this project is to build a fundamental frequency estimator for singing voices, which have a much wider pitch range than speech signals. If we apply SRPD directly, it will take too much computation to perform the auto-correlation at every possible fundamental frequency candidates. Hence, the idea is to apply the maximum likelihood harmonic matching method to get a first guess on the fundamental frequency in a very rough precision. The SRPD method is then applied to obtain higher resolution estimates. This hybrid method was tested to retrieve the fundamental frequency time trajectories from the coloratura-soprano recordings, a glissando (from E6 (1318.5 Hz) to F3 (174.61 Hz)) vowel /a/ and a sustained C5 vowel /o/ tone. Results show that the method is effective.

2. Method

The proposed fundamental frequency estimation method consists two phases - maximum likelihood harmonic matching and SRPD. The input signal is decomposed into overlapped windowed frames. In the first phase, the fundamental frequency of each frame is detected in half-tone precision for the coloratura-soprano range from note F3 (174.61 Hz) to F6 (1396.9 Hz). In phase II, the SRPD is then applied to refine the fundamental frequency estimates trajectory. The frequency resolution of the fundamental frequency estimates is infinite via SRPD. The time resolution is 5.8 ms, determined from the hop size of the window frames. Hence, the hybrid fundamental frequency estimator can track the time variation of the fundamental frequency up to 172 Hz.

2.1 Front end processing

The front-end features are the measured partials defined as the peaks of the amplitude of the short-time Fourier transform (STFT) spectrum.

Step 1: The input signal is decomposed into overlapped frames with hop size 5.8 ms. The hop size is the time interval between the centers of two adjacent frames. Each frame is windowed by a 46 ms long hamming window. The hamming window is chosen since it is the typical window used for speech signals. The window length must be chosen to trade off the reduced time resolution available when using a long window with reduced frequency resolution available with a short window. In order to resolve reliably the partials of a harmonic signal with fundamental frequency of f_0 Hz, it is necessary for the spectral bandwidth of the analysis window to be less than $f_0/2$ Hz, corresponding to a hamming window about four cycles long. Thus, the 46 ms window length is satisfactory for f_0 above 87 Hz [4].

Step 2: Perform the short-time Fourier transform on each frame.

Step 3: In this step, we want to extract the partials of each frame [2]. In order to find the candidate partials, we can consider that the spectrum values around a maximum are represented by a short-term spectrum of a sinusoid at the same frequency. Let Φ_i be the candidate partial, S the observed spectrum from *Step 2* and W_{Φ_i} the spectrum of a windowed sinusoid at frequency Φ_i. The confidence of the candidate partials can be measured by the normalized cross correlation between the observed spectrum and the spectrum of the windowed sinusoid, $\langle S, W_{\Phi_i} \rangle / \|S\| \cdot \|W_{\Phi_i}\|$. We only consider candidate partials with confidence level greater than 0.1. The remaining candidate partials are then sorted by the confidence measures. The first 20 candidate partials are stored as the measured partials for this frame. The number of measured partials is often less than 20 for our available recordings. Figure. 1 shows the partial time trajectories of part of the training set. The upper 8 plots are the partial trajectories of vowel /a/ from note F4 to F5 (from left to right). The lower 8 plots are the ones of vowel /e/.

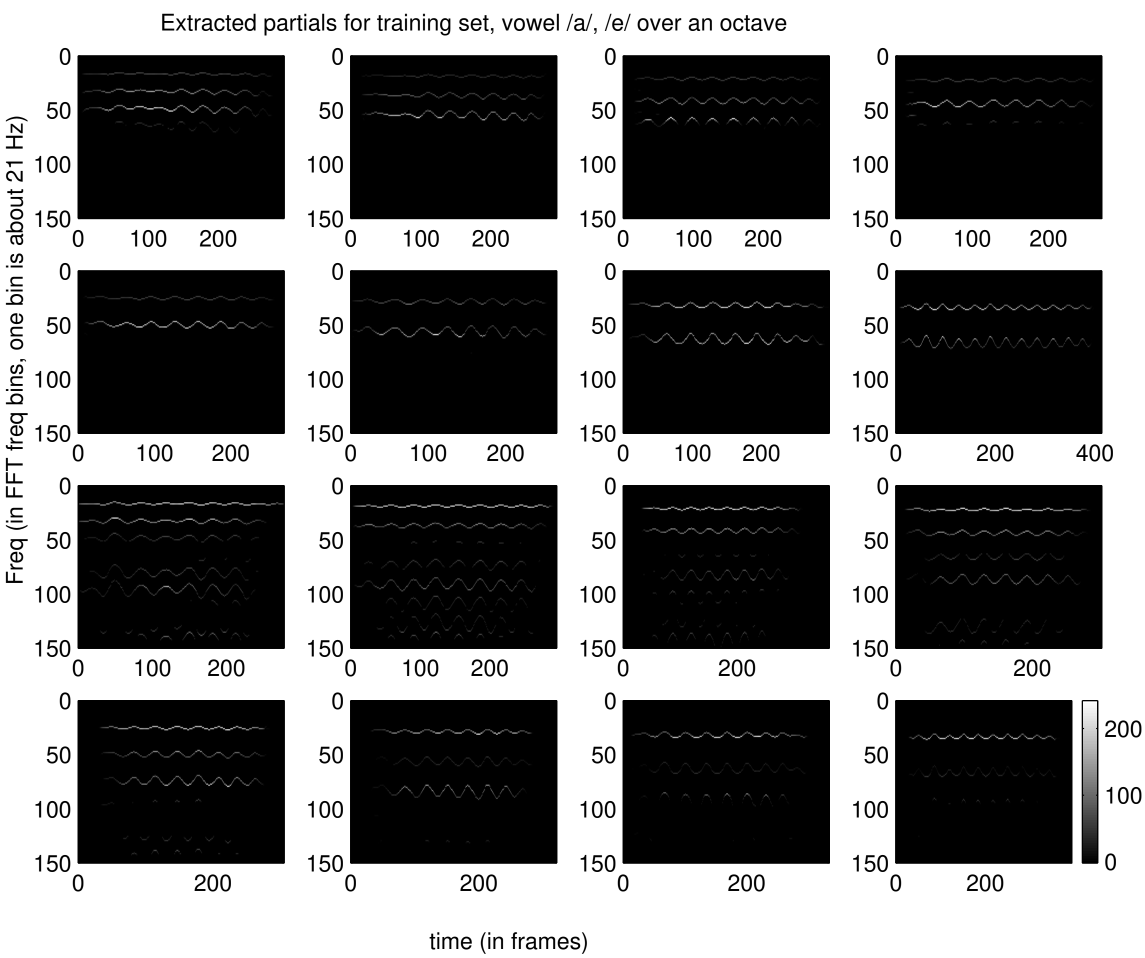
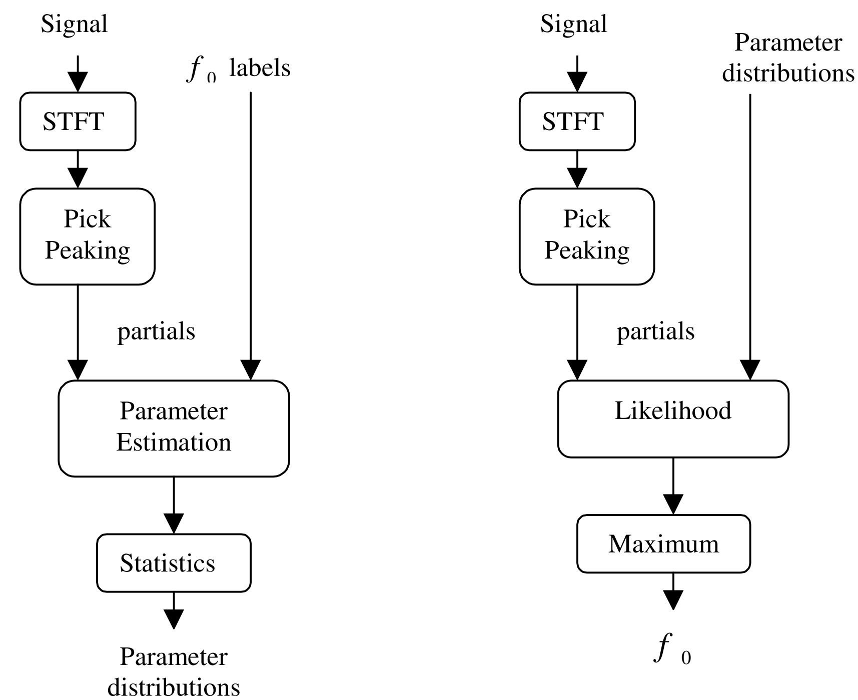


Figure. 1 The partial trajectories of some training set.

2.2 Phase I: maximum likelihood harmonic matching method

This method [1] [2] [3] considers the fundamental frequency estimation as a spectral pattern recognition problem. The original method developed by Doval B. is meant for general musical sounds. By learning the distribution of spectral parameters like frequency or the amplitude of the partials, the fundamental frequency is estimated via maximum likelihood. This method can be illustrated by the following flowcharts. The left chart shows the block diagram of the learning stage and the right chart shows the block diagram of the recognition stage.



The drawback of this method is that one can not obtain infinite resolution for the fundamental frequency estimate with a fix time resolution (the trade-off of time resolution and frequency resolution is described in Section 2.1). However, it is good enough for our requirement, half-tone resolution for the fundamental frequency estimates.

To reduce the computation, We have simplified the algorithm toward singing voice applications. The following paragraphs describe the details of the estimation procedure.

Denote the observation O as the set { m_i } of so-called partials found in the STFT. An observation contains harmonic partials of the f_0 being sought. If the frequency of the observed partials belongs to the interval $[(k-0.5)f_0,(k+0.5)f_0]$, this partial is called the kth harmonic partial. Let $\{mh_k\}_{k\in H}$ be these harmonic partials, where H is the set of k for which the kth harmonic is present in O. Some harmonics of the target f_0 may be absent and not present in O, let $\underline{H}$ be the set of k for which the kth harmonic is absent. Because of the –12dB to –18dB roll-off of the glottal signal spectrum, the partials at high frequency range of the voice signals may be too small to detect. Hence, some of the higher order harmonic partials are not detected when f_0 is high. The sets H and $\underline{H}$ provide cues for true f_0. Let $p_H(k)$ denote the probability of the presence of the kth harmonic and $p_{\underline{H}}(k)$ denote the probability of the absence of kth harmonic, where $p_H(k)+p_{\underline{H}}(k)=1$.

The voice signal is not purely periodic; hence, the measured partials are not exact integer multiples of f_0. Moreover, vibrato, frequency deviation from the fixed note, is always present in the singing voice. To account for the frequency deviations from the labeled pitch note, the frequency distribution around the kth harmonic is learned. Let $f_{mh_k}(\cdot)$ denote the probability density value of the kth harmonic partials. The likelihood of f_0 estimate is then defined as $\prod_{\underline{H}} p_{\underline{H}}(k)\cdot\prod_{\underline{H}} p_H(k)\cdot\prod_{\underline{H}} f_{mh_k}(\frac{m_{i(k)}-k\cdot f_0}{k\cdot f_0})$. Where $m_{i(k)}$ is the frequency value of the kth harmonic partials.

Learning stage: In this stage, the probability distribution $p_H(k)$ and $f_{mh_k}(m_{i(k)})$ is learned. The training data are 40 coloratura-soprano recordings. They are sustained vowels, /a/ /e/ /i/ /o/ /u/, from note F4 to F5 with whole-note increment. The whole data set is used to count the number of kth harmonic partials present so that $p_H(k)$ can be calculated. The distribution $f_{mh_k}(\cdot)$ is learned for each kth harmonic and for each whole note from F4 to F5 (There is no significant difference if we just use one distribution for every note). The Gaussian distribution is used to describe the distribution $f_{mh_k}(\cdot)$ for simplicity. The sample mean, $\hat{\mu} = \frac{1}{N_k}\sum_{n=1}^{N_k} x_n$ and variance, $\sigma^2 = \frac{1}{N_k}\sum_{n=1}^{N_k}(x_n - \hat{\mu})^2$, (the biased variance estimate is used here) are the estimated parameters for the Gaussian distribution. Where N_k is the total number of kth harmonic partials of the training data and x is $(f - k \cdot f_0)/(k \cdot f_0)$ with f, the frequency of the kth harmonic partial.

Recognition stage: After obtaining the harmonic partials, the likelihood is calculated for each possible note from F3 to F6 with half-tone increment. The estimated note is the one with maximum likelihood.

2.3 Phase II: SRPD (super resolution pitch determination method)

The idea is that the cross-correlation of two adjacent segments of a periodic signal is maximized when they are spaced apart by a period or multiple of the period [5]. Since we have refined the fundamental frequency search range in phase I, the ambiguity of multiple periods are resolved. This intuition has been derived via Bayesian approach [7] (more precisely, we should say maximum likelihood from the following derivation). We will set off from this point of view.

Denote the voice signal as $s(t)$. By the quasi-periodicity assumption, it follows that $s(t) \approx a \cdot s(t+T)$, where T is the period and a is the inevitable amplitude scaling. Considering two segments of $s(t)$ with d samples and sampling period τ, one could define vectors $v(t) = [s(t) \quad s(t+\tau) \quad ,..., \quad s(t+(d-1)\tau]$ and $v(t+T')$, which is T' times the lag of $v(t)$. Since the signal is quasi-periodic, we can model the difference between $v(t)$ and $a \cdot v(t+T)$ as a zero mean Gaussian distribution. Let us assume that the covariance matrix is diagonal and with equal standard deviation, σ. H_T denotes the hypothesis that the period is T. We can then write down the probability density function as

$$p_{v|H_T}(v(t) \mid H_T) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{\|v(t) - a(t,T)v(t+T)\|^2}{2\sigma^2}) \qquad (1).$$

Where $a(t+T)$ can be obtained by minimizing the exponent, hence,

$$a(t+T) = \frac{v(t)^T v(t+T)}{\|v(t+T)\|^2} \qquad (2).$$

The best estimate for the pseudo-period is simply the choice of the hypothesis that will maximize the *a posteriori* probability density function, $p_{H_T|v}(H_T \mid v(t))$. Using Bayes' rule, the goal is to find T which will maximize

$$p_{H_T|v}(H_T \mid v(t)) = \frac{p_{v|H_T}(v(t) \mid H_T) \cdot p_{H_T}(H_T)}{p_v(v(t))} \qquad (3),$$

where $p_{H_T}(H_T)$ is the *a priori* probability density function that T is the period of the signal, and $p_v(v(t))$ is the probability density function for $v(t)$ and assuming an equal *a priori* probability that the signal has a fundamental period T across all possible periods and recognizing that $p_v(v(t))$ is independent of T. Maximizing equation (3) is equivalent to maximizing equation (1). This is equivalent to finding period T that minimizes the exponent of (1). Plugging (2) into (1), we can derive that the T estimate needs to maximize $\frac{v(t)^T \cdot v(t+T)}{\|v(t)\| \cdot \|v(t+T)\|}$, which is the normalized cross-correlation between $v(t)$ and $v(t+T)$.

The estimated T from the previous derivation has a resolution limited to the sampling rate of the input signal. One can further refine the resolution by interpolation. The details can be found in the original SRPD paper [5]. This interpolation step is not essential for our recordings that have a sampling rate of 44100 Hz.

3. Results and Conclusion

Two recordings are tested for this hybrid fundamental frequency estimator. One is a coloratura-soprano glissando recording with pitch range from E6 (1318.5 Hz) to F3 (174.61 Hz)). Although the pitch range is out of the range (F4 to F5) that we trained the parameter distribution of the model, it can still get decent estimates. Figure. 2 shows the estimated fundamental frequency trajectories over time frames in the upper plot and the associated log likelihood function in the lower plot. The solid line in the upper plot is the final result from the hybrid estimator and the dotted line is the rough estimate from phase I. The likelihood function also provides a confidence measure of the estimate. At the unvoiced segments (the beginning and the end of the recording), the likelihood function gives very low values.

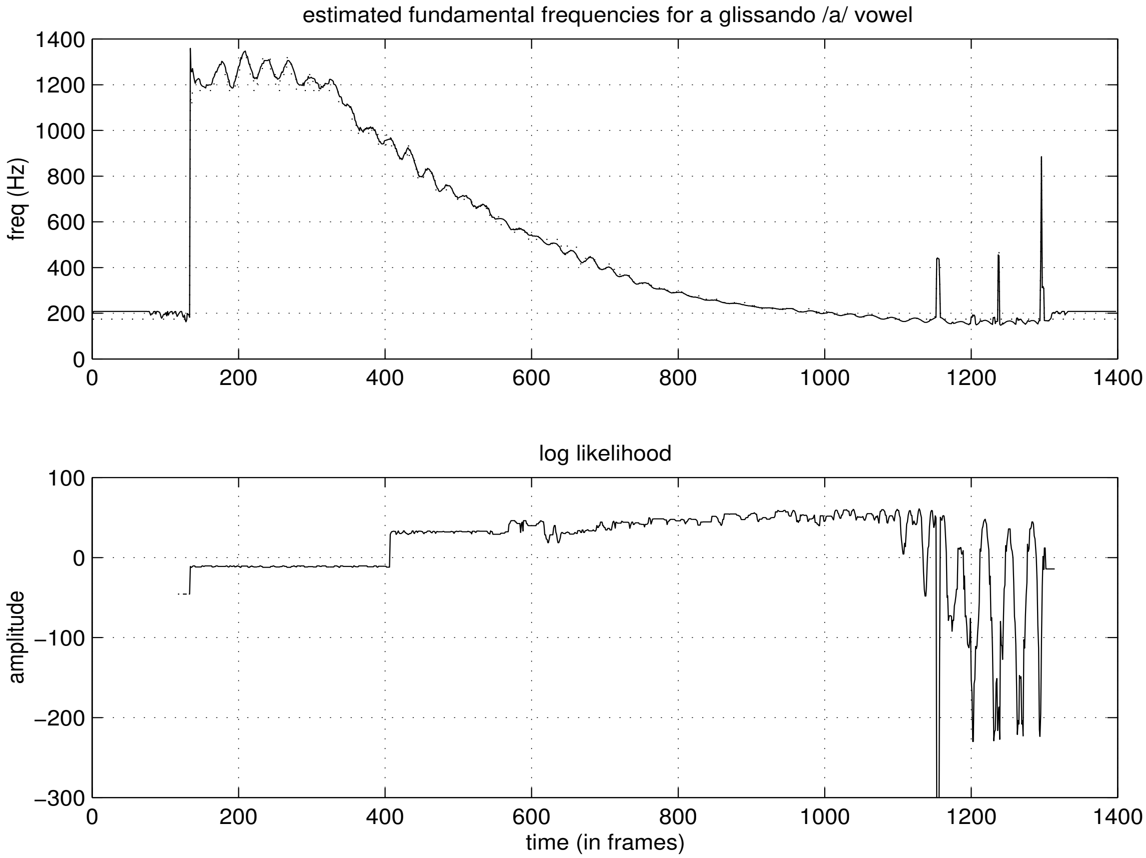


Fig. 2 The result for the glissando recording.

In order to verify the accuracy of the estimate, the synchronized differentiated EGG (DEGG) signal is also obtained from the midiVox device [6]. Using the DEGG signal, we obtain another fundamental frequency estimates trajectory in cycles of the signal. The test recording is a C5 vowel /o/. Figure. 3 shows the log likelihood function for the hybrid method, the results of the estimated fundamental frequency trajectory from the hybrid method, and the measured fundamental trajectory from the DEGG signal. Since the time resolution of the estimated and measured fundamental frequency trajectories are different, these two trajectories can not be overlapped and compared.

However, they show the same vibrato depth and the same number of vibrato cycles. This shows that the hybrid method is effective.

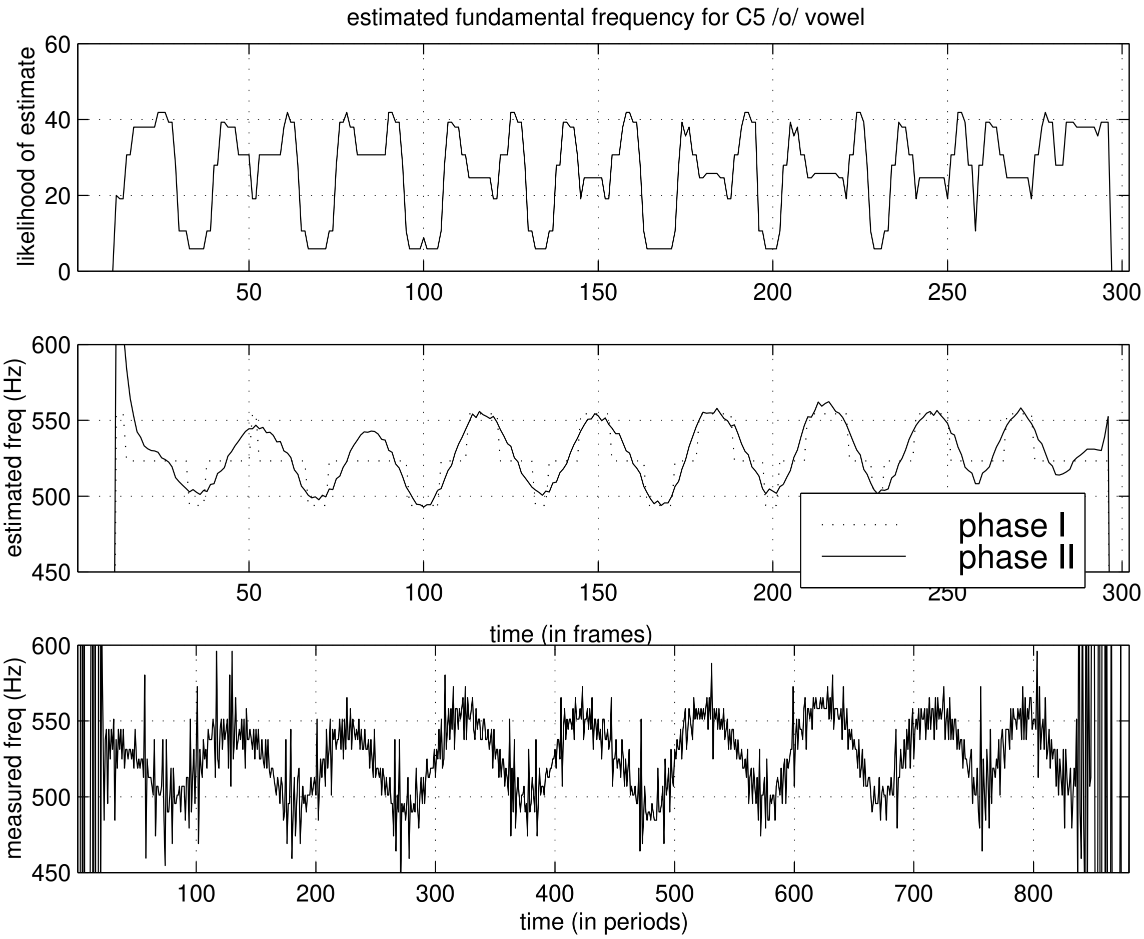


Fig. 3 The result for the sustained vowel /o/ at C5 note.

From the result of the glissando recording, we can see the estimation discontinuities at the end. This can be improved by applying the continuity constraint on the estimates from phase I. One could use an HMM model to model the transition probability between different notes suggested by Doval B. [2] or broaden the search range of phase II when the likelihood from phase I is low.

Reference

[1] Doval B., "Fundamental frequency estimation of musical sounds using statistical learning," *ISMA, pp. 454-457,* 1995.

[2] Doval B., X. Rodet, "Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs," *ICASSP*, Vol. 1, pp. 221-224, Minneapolis, April 26-30, 1993.

[3] Doval B., Rodet X., "Fundamental frequency estimation using a new harmonic matching method*," ICMC (International Computer Music Conference)*, pp. 555-558, Montreal, Oct. 16-20, 1991.

[4] R. C. Maher, J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure,", *J. Acoust. Soc. Am*. 95(4), April 1994.

[5] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. on Signal Processing*, Vol. 39, No. 1, pp. 40-47, Jan. 1991.

[6] http://www.synthony.com/vintage/midivox.html, SYNCHROVOICE, INC.

[7] W. D. Oliver*,* http://theremin.media.mit.edu/~woliver/singhtml/.