A Pitch Extraction System Based on Phase Locked Loops and Consensus Decision

Patricia A. Pelle, Claudio F. Estienne

School of Engineering, University of Buenos Aires, Argentina
ppelle@fi.uba.ar, cestien@fi.uba.ar

Abstract

In this work a very low error rate pitch estimation system is presented, which is also very robust against noise. Two key aspects of the system are mainly responsible for such good behavior: on the one hand we use a multiple estimation scheme based on PLL’s. These devices provide us with robust information about the period of the speech signal harmonics. By combining this information with an additional independent estimation it is possible to obtain a robust estimation of \( f_0 \). On the other hand, multiple estimations are combined in a stage that assesses each of them, retaining the more reliable ones. A final agreement value between these qualified estimations is the final result of the system. This consensus decision significantly improves the initial estimation accuracy. Overall performance is assessed by comparing our system to the get\_f0 algorithm, under clean and noisy conditions. We show that our system outperforms get\_f0 over all presented conditions.

Index Terms: pitch estimation, consensus decisions, PLL frequency estimation, PLL noise robustness.

1. Introduction

Pitch detection is a complex problem. Many difficulties arise when estimating pitch including pitch-doubling, pitch-halving, performance degradation with noise, voiced-voiced classification and estimation at the beginning and end of voiced segments. These factors make pitch detection and extraction a difficult task, and various algorithms have been reported in the past [1] [2] [3]. Some of them are reliable for limited applications, but nearly all fail in noisy environment conditions. Reliable pitch detection is important in determining prosodic features like stress, rhythm, and intonation. It is also important in distinguishing segmental categories in tonal languages, speech coding systems, and speech analysis-synthesis systems [1].

In this work an approach to pitch detection is presented which combines two factors to give a very good performance for pitch estimation. First, the use of Phase Locked Loop (PLL) devices give several primary estimations of the pitch frequency. PLLs are widely used in communications systems, including FM demodulation, frequency multiplexing, and frequency synthesizers [4]. These devices have interesting properties, such as the ability to automatically track periodic signals, and extract their instantaneous phase under severe noise conditions. Those characteristics motivated us to explore the use of PLLs for speech feature extraction [5] as well as pitch estimation [6, 7], where PLL robustness against noise is exploited to find robust estimates of pitch. Using multiple PLLs that estimate the frequency of portions of the signal spectrum, combined with auxiliary estimations of \( f_0 \), a set of estimations of pitch frequency are obtained for each frame of signal. The second factor is that these estimations are considered hypotheses for the true value of \( f_0 \) and are combined with other side information to obtain a unique final value of \( f_0 \). Side information allow us to classify the hypotheses, and to prune the set to retain only those considered sufficiently good. The final value will be the median value of the pruned set, which can be considered an agreement value within the set of good hypotheses. This kind of consensus decision between reliable values can overcome individual errors resulting from the hypothesis generation stage, and give a very good overall performance for the system.

The rest of the work is divided as follows: In section 2 we describe our pitch detection system and the function of each block; in section 3 more details about the blocks that construct the system are described. Section 4 states experiments performed and databases used with the results obtained, and in section 5 we discuss experimental results giving some concluding remarks.

2. Overall system description

From a functional point of view the system is composed of three subsystems (Fig. 1). The goal of the first subsystem is to obtain a set of hypotheses for the \( f_0 \) value in each frame. Such hypotheses are pitch estimations based on the frequencies of the speech signal harmonics. This subsystem, which will be explained in details later, can be briefly described as a band pass filter bank followed by PLLs appropriately adjusted in order to give an indication of the signal frequency present in the frequency range of the filter. For each filter-PLL block, a complementary coarse estimation of pitch frequency is made, which, combined with the PLL frequency indication, gives a more accurate estimation of \( f_0 \) than the coarse one. The next subsystem is devoted to assess the accuracy of each hypothesis generated in the first subsystem. Side information calculated during the hypothesis estimation step is taken into account in this subsystem and used to calculate three scores that indicate whether the hypotheses should be accepted or not. Hypotheses that exceed the established thresholds are allowed to be passed to the final subsystem while the rest are discarded. The last subsystem has a buffer that contains hypotheses approved in the previous stage for three consecutive frames, and outputs the median of these values as the final pitch estimation of the overall system. This approach constitutes an extension of a median filter, which is commonly used in pitch detection systems in general, but in this case is applied to several values per frame.

It should be noted that once the hypotheses are generated, other approximations may be implemented to chose the best value from them. But this approach, which implements a kind of consensus between the approved hypotheses, has shown a great reduction in the error rate. The error rate of the set of hypotheses generated in the first subsystem may be reduced by a
factor of up to four times in this way. An explanation for this behavior is that even though internal components may make mistakes, as long as the majority of the hypothesis is correct, their median value will be correct.

3. Detailed subsystems description

3.1. Hypotheses generator system

The \( f_0 \) hypotheses generator is based on different measurements of the frequency of the signal harmonics, and it is similar to the pitch estimation presented in a previous work [7]. The main idea is to use information about the frequency of the signal harmonics, combining them with an additional coarse estimation of the fundamental frequency to complete the \( f_0 \) estimation. This stage can be described as the combination of three different processes:

3.1.1. Filterbank and PLL’s processing

Information about the frequency of different harmonics can be obtained with a band pass filter bank, and a PLL following each filter, in an arrangement as shown in Fig. 2. The goal of the filters is to select the frequency range that each PLL should be able to synchronize with. Each PLL is set to a free running frequency equal to the cutoff frequency of the corresponding filter. The number of filters was determined experimentally in order to cover up to the maximum pitch frequency that the system can detect. We used 20 channels linearly spaced in Mel scale and approximately a constant \( Q \) factor covering the range between 50Hz and 5000Hz. We also found that band pass filters with an asymmetrical frequency response perform better than those with symmetrical response. We have chosen the kind of filters suggested by Wang and Shamma [8] and we empirically adjusted the \( Q \) factor and the degree of asymmetry of the filters.

This pitch estimation is based on the fact that it is straightforward for a PLL to obtain a highly precise estimation of the frequency of a periodic signal. The PLL is a nonlinear device that makes use of a feedback closed loop to set an internal sinusoidal oscillator to be locked in phase with the instantaneous phase of the input signal, and follows its fluctuation in a very precise and robust way. A filter in the path of the signal loop is responsible for the good behavior under noisy conditions, and also for the dynamics of the PLL. If the internal oscillator and the signal are in lock, the frequency estimation is very reliable. An additional signal, called the lock indicator, is generated in the PLL to indicate whether or not the PLL is in lock. This signal is one of the side signals used in the selector system. More details concerning analog PLLs design can be found in [4].

3.1.2. Coarse fundamental frequency estimation

Estimation of the coarse fundamental frequency \( \hat{f}_0 \) is accomplished by calculating the Discrete Fourier Transform (DFT) of a section of the input signal \( s(t) \), called \( s_{\omega}(t) \). If the spectrum of \( s_{\omega}(t) \) is sharp, with big differences between peaks and valleys, it is possible to obtain the desired estimation by performing the inverse DFT of the spectrum modulus, in our case band limiting it up to 2000Hz. By doing this, a time signal is obtained which has its highest peaks in multiples of the fundamental period \( T_0 = 1/f_0 \). The coarse estimated frequency is measured in the resulting signal as the inverse of the time interval between zero and the occurrence of the highest peak.

One key component of this simple estimation procedure is the construction of a very sharp spectrum of the input signal. One way to do this is to increase the length of the window section, but long windows would not allow us to follow fast changes in the pitch of the signal. Our criteria to select a suitable window section without increasing unnecessarily the window width is to take a segment length that covers a number of periods greater than two. Sections of the input signal of more than this width possess a spectrum with valleys between the peaks corresponding to the position of each harmonic. We select a window width based on the period of the signal read by the PLL. As we do not know to which harmonic number \( f_{\text{pll}} \) corresponds to, we choose a uniform harmonic number equal to 6, which is a number sufficiently high to show experimentally good behavior.

Information of the frequency \( f_{\text{pll}} \) may be used in another way in this estimation in order to improve the \( f_0 \) accuracy. The IDFT of spectrum signal may contain many peaks, an determination of the highest peak is a frequent source of gross errors. But, considering that \( f_{\text{pll}} \) is a harmonic frequency, i.e. \( f_{\text{pll}} = f_0 k \), or equivalently that \( T_{\text{pll}} = 1/f_{\text{pll}} = T_0/k \), we conclude that we only need to find \( k \) peaks, each of them positioned in intervals centered in multiples of \( T_{\text{pll}} = 1/f_{\text{pll}} \). Then, the determination of the highest peak is restricted to choosing the highest of the \( k \) peaks determined in intervals of the PLL period. This restriction over the search of the highest peak produces better performance in the \( f_0 \) determination. And, as for the width of the window signal \( s_{\omega}(t) \), \( k = 6 \) results in good performance.

3.1.3. Combination of coarse and PLL-filterbank estimations

Finally it is possible to combine the coarse estimation with the very precise PLL frequency harmonic indication to obtain a more accurate estimated \( f_0 \). The fact that the PLL frequency indication, \( f_{\text{pll}} \), was a harmonic of the fundamental frequency, may be expressed as

\[
f_{\text{pll}} = f_0 k,
\]

where \( k \) is an unknown integer. But using the additional coarse estimation of the pitch frequency, \( \hat{f}_0 \), it is possible to estimate this number \( k \) as the closest integer to \( f_{\text{pll}} \) and \( f_0 \) ratio, and then
\( f_0 \) will be calculated as
\[
\hat{f}_0 = f_{\text{pll}}/\text{round}\left( f_{\text{pll}}/\hat{f}_0 \right)
\]

3.2. Hypothesis evaluation and selector

As was mentioned previously, the goal of this subsystem is to produce a pruning of the hypotheses set, retaining only those considered sufficiently good. This goal is accomplished in two steps. In the first step, a subset of the hypotheses corresponding to the higher lock indicator signal is chosen. In this way some hypotheses that are clearly incorrect because the \( f_{\text{pll}} \) indication has no relation with the input signal are discarded, because a low lock indication denotes that the corresponding PLL is not locked to a periodic signal. In the next step three scores (described below) are calculated for each and compared with corresponding thresholds. Threshold setting is a key feature that has great importance in the overall system performance. It is desirable to establish a threshold that is neither too demanding to retain any hypothesis, nor too indulgent as to approve too many hypotheses that are not correct (false positives). In this work, thresholds are set experimentally to obtain the maximum number of sets of three frames hypotheses with more true positives than false positives, but with at least one hypothesis remaining in the set.

The scores that we use intended to carry out indications of problems or anomalies in the the pitch frequency estimation.

3.2.1. score 1: Geometric considerations

This score measures the agreement between \( f_{\text{pll}} \) and \( f_0 \), by calculating the ratio of both frequencies apart from an integer:
\[
\text{score1} = \left| f_{\text{pll}}/f_0 - \text{round}(f_{\text{pll}}/f_0) \right|
\]

Values near 0 are indicating that \( k = \text{round}(f_{\text{pll}}/f_0) \) is probably well estimated with these frequencies.

3.2.2. score 2: Position of the highest \( \hat{f}_0 \) peak

When \( \hat{f}_0 \) is calculated, a critical fact is that the highest peak does not always correspond to the signal period. It could, for example, appear at \( 2\hat{f}_0 \). This is a clue that the estimation is likely to be incorrect. In the normal case, a second highest peak occurs at two times the first one. So the score will be calculated as
\[
\text{score2} = \left| 1 - \text{peak}_1/\left(\text{peak}_2 - \text{peak}_1\right) \right|
\]

Values near 0 express a good behavior of the peaks, and probably a good estimation of \( f_0 \), while an interleaved peak position corresponds to a value of 3.

3.2.3. score 3: \( f_{\text{pll}} \)-lockin\(_{\text{pll}} \) distribution

Unlike the other two scores, the last score calculation implements a comparison between the distribution of PLL frequencies and the estimated \( f_0 \) value. If \( f_0 \) was a good estimation, we would expect any \( f_{\text{pll}} \) to be positioned around \( f_0, 2f_0, 3f_0 \), etc, at least for those frequencies that have high associated lockin. So we calculate \( \text{score3} \) as
\[
\text{score3} = \left( \sum \text{lockin}_i \cos(2\pi f_{\text{pll}}/f_0) \right)^2 / \sum \text{lockin}_i^2
\]

where if \( f_{\text{pll}} = k_i f_0 \), \( i \), the calculated score is maximum.

4. Experiments and results

4.1. Experiments and data description

Performance is evaluated using Keele pitch extraction reference database [9]. Pitch reference is provided from simultaneously recorded laryngograph trace. The database is available at \(<\text{ftp.cs.keele.ac.uk/pub/pitch}>\). It consists of five male and five female speakers, each speaking a short story of about 35 seconds. The Keele database is studio quality, sampled at 20 KHz.

Results are compared to those of \( \text{get}_0 \) algorithm [10], a well known pitch extraction algorithm available as part of Wavesurfer toolkit (see \(<\text{http://www.speech.kth.se/wavesurfer}>>\)). The frame rate is set to 10 msec and the range for frequency estimation from 50 to 500Hz in both our system and Wavesurfer. Other parameters of Wavesurfer are left at their default values.

Accuracy was evaluated in terms of gross error rate (GER), measured as the percentage of frames in which estimated frequency deviates from the reference by more than a certain amount (20% in our case). Furthermore, for voiced frames in which the error is smaller than 20%, we evaluated fine errors measured by the mean and standard deviation of the absolute error. For purpose of comparison with other results on pitch detection ([13], [11], [12]), we divided voiced frames into two sets based on pitch estimated from Wavesurfer: “clearly voiced frames”, where reference voiced segments are detected as voiced by Wavesurfer, and “voiced-to-unvoiced frames”, where reference voiced segments are wrongly detected as unvoiced.

For comparison purposes, we considered as gross errors those voiced frames detected as unvoiced by Wavesurfer.

4.2. Results

Table 1 shows the performance of our system under clean conditions compared against that of the previous version of this system [7] and \( \text{get}_0 \) performance. We show accuracy in terms of GER, mean absolute error (MAE), and standard deviation of the absolute error (STD), both in the “clearly voiced” set of frames case and in the whole set of voiced frames case (clearly voiced plus voiced-to-unvoiced frames).

Table 2 shows performance of the system under babble noise conditions. As noise is increasing, portions of “clearly voiced signal” reduce in length, and the GER in this portion of signal tends to be very low. Accuracy is measured with the addition of noise to the signal resulting in SNRs from 30dB to 0dB, in 10dB steps. Noises are taken from NOISEX database examples that are available on the Rice University Digital Signal Processing (DSP) group home page, at \(<\text{http://spib.rice.edu/spib/select_noise.html}>\).

5. Summary and discussion

Table 1 shows that the proposed system outperforms the other two systems for both performance measures and both sets of frames, the clearly detected voiced frames and the actual set of voiced frames. Furthermore, we can see that the performance of the proposed algorithm degrades significantly less than that of wavesurfer when considering the actual set of voiced frames instead of the correctly detected set. An advantage over the previously presented method (without scoring) can also be observed. This result can be predicted by the lock-in property of PLLs on which our system is based. Whenever an unvoiced to voiced transition occurs in the speech signal, PLLs of the bank
will tend to lock to various harmonics of the signal, and nearly immediately an estimation of the fundamental frequency will be present at the output of the system. This fact makes our system intrinsically more efficient in the detection of voiced frames. As a consequence gross errors, which are more likely to occur at the beginning and end of voiced sections, will also be lower. Also it is important to note the benefits obtained by adding the selection stage, which further improves the performance of the system.

Table 2 shows that the PLL based system outperforms the Wavesurfer algorithm both for gross and fine errors, and both in clearly voiced frames chosen by Wavesurfer estimation as well as in the actual set of voiced frames for all noise levels. This is also a consequence of the PLL behavior under noisy conditions. As mentioned before, PLLs lockin property makes them highly immune to noisy conditions. If the PLL filter bank is still capable of locking to the signal frequency, the rest of the system will consequently be unaffected by the noise and no significant degradation of the performance will occur. Furthermore, a better behavior under noise is achieved from the use of the median filter of the best scored detections in the last stage of the system, which allows for individual errors to be ignored, not affecting the overall performance of the system.

Finally, we believe that future work can lead to further improvements in this system. For example, it is no difficult to imagine that a reconstruction of the first harmonic of the signal could be obtained from other internal signals in the PLL’s devices not considered here, providing us with important synchrony information about the pitch signal.

### 6. References


