A ROBUST PITCH EXTRACTION SYSTEM BASED ON PHASE LOCKED LOOPS

Patricia A. Pelle

University of Buenos Aires, School of Engineering

ppelle@fi.uba.ar

ABSTRACT

We propose a new scheme for pitch estimation using phase locked loops (PLL). These devices possess desired properties for speech signal analysis, including instantaneous phase detection, and noise robustness. The method is based on a PLL arrange applied to the outputs of a bank of passband filters, each of them tuned to a portion of the spectrum of the signal. PLL devices provide us with robust information about the period of the speech signal harmonics. Combining this information with the spectrum of selected portions of the signal, it is possible to determine a reliable estimate of the period of the whole signal. Performance is evaluated by comparing our system to get_f0 algorithm, with various noise levels. We show that our system largely outperforms get_f0 under noisy speech conditions.

1. INTRODUCTION

Pitch detection is a complex problem. Many difficulties arise when estimating pitch including pitch-doubling, pitch-halving, performance degradation with noise, voiced-unvoiced decision and estimation at the beginning and ending 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 system, and speech analysis-synthesis systems [1].

In this work we present a novel pitch detection system based on phase locked loop (PLL) devices. PLLs are widely used in communications systems, including FM demodulation, frequency multiplexing, and frequency synthesizers [4]. PLLs have interesting properties, as the ability to automatically track periodic signals, and extracting their instantaneous phase also under severe noise conditions. Those characteristics motivated us to explore the use of PLLs for speech features extraction [5] as well as pitch estimation [6].

In the present work we outperform the results presented in [6] by using many harmonics of the speech signal to extract pitch information rather than utilize the first one only. We also include phase information provided by the PLLs to take a window of a suitable number of periods of the speech signal, in order to make a rough estimation of the fundamental frequency. This estimation which is not necessarily very precise, is used to obtain an indication of the approximated fundamental frequency. Combining information on the frequency of the best estimated harmonic with this approximated fundamental frequency we obtain very low gross error estimates as well as a good behavior of fine errors.

We tested the performance of our system under both clean and noisy speech. We found that our system performs nearly equal to traditional pitch estimators in clean conditions, but it highly outperforms them under various levels of noise added.

The rest of the work is divided as follows: In section 2 we outline the operation principles of PLL devices; in section 3 we describe our pitch detection algorithm; in section 4 we show and discuss experimental results giving some concluding remarks.

2. PHASE LOCKED LOOP OPERATION

A PLL consists on a loop containing three basic blocks [4] (Fig.1): a voltage-controlled-oscillator (VCO) whose frequency is controlled by an external voltage, a phase detector which is usually a multiplier, and a low-pass filter (Loop-filter). The phase detector compares the phase of a periodic input signal against the phase of the VCO output resulting in an error signal which is a function of the difference between instantaneous phases of the input ($\theta_i(t)$) and VCO ($\theta_o(t)$). This error signal is then filtered and amplified by the loop filter, and applied as a control voltage to the VCO. The VCO output is fed then as input to the phase detector. The VCO operates at a set frequency known as free-running frequency (ω_0). The control voltage forces the output frequency of the VCO to vary in a direction that reduces the phase difference between VCO output and the input signal. If both phases are sufficiently close, negative feedback makes the VCO to lock or synchronize with the incoming signal. Once in lock, both VCO output and input phases are identical, and as a consequence, their frequencies are also equal.

When the incoming signal is poly-harmonic, the lock-in condition depends on the VCO free-running frequency, and

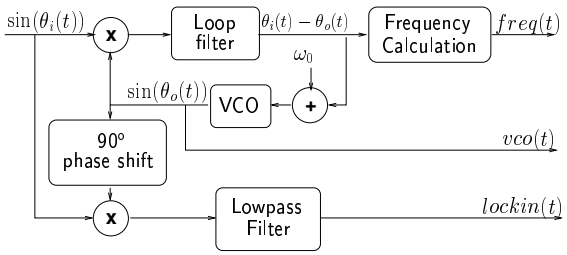

Fig. 1. Basic PLL operation.

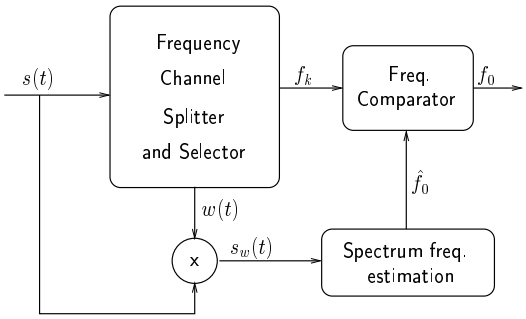

Fig. 2. Overall system diagram.

on the relative energy of signal harmonics. Also it depends on the loop filter parameters which make PLL to be able to *capture* a frequency. Higher energy harmonics will have more chances to be synchronized with a PLL than lower energy ones.

In our scheme we also provide an indication that the PLL is locked. It is generated with a quadrature phase detector followed by a smoothing filter. When the main phase detector output tends to zero (locked condition), the output of the second phase detector tends to be maximum, and a measure of the locking degree of the main loop is obtained. The smoothing filter is necessary to avoid flickering of the lock indicator signal. This signal is not only useful in the detection of voiced-unvoiced segments for speech signals, but also can be seen as a measure of the reliability of frequency and phase indications of the PLL.

For the rest of the work we will consider a PLL as a block with one input signal and three output signals: Frequency signal $freq(t)$ that indicates the instantaneous frequency at which the PLL is locked; VCO output $vco(t)$ that is a sinusoidal signal locked in phase with the input signal; and lock indicator signal $lockin(t)$ that as said is an instantaneous measure of the degree of lock of the PLL. We implemented an algorithmic version of the analog PLL, further details concerning analog PLLs design can be found in [4].

3. PITCH ESTIMATION SYSTEM

3.1. System overview

The proposed system is based on the fact that it is straightforward for a PLL to obtain a highly precise estimation of the frequency of some harmonic of a periodic signal. We could also roughly force the PLL to lock to a desired harmonic (or a range of candidate harmonics) by band-pass filtering the periodic signal. On the other hand if we had at least a coarse estimation of the fundamental frequency f_0, we would be able to determine the number of the harmonic to which the PLL locked simply dividing the PLL frequency by the estimated fundamental frequency. Finally we could combine coarse estimation of f_0 and precise estimation of the harmonic to obtain a more accurate estimation of f_0. The operation of our system is described by three main blocks as shown in Fig. 2.

The frequency channel splitter and selector block is devoted to obtain a precise estimation of a harmonic of the fundamental frequency f_k of a speech signal $s(t)$. This block also provides a suitable signal $w(t)$ for windowing $s(t)$, necessary to obtain $\hat{f}_0$, which is a coarse but simple estimation of the fundamental frequency. The spectrum frequency estimation block obtains this coarse estimation, and finally both estimation f_k and $\hat{f}_0$ are combined in the frequency comparator block resulting f_0 the overall estimation of the fundamental frequency.

3.2. Frequency channel splitter and selector

This stage consists on a band pass filter bank, each followed by a PLL arrangement (Fig. 3). The goal of the filters is to select the range of frequencies that each PLL would be able to synchronize. 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 $500Hz$. 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 [7] and we adjusted empirically the Q factor and the degree of asymmetry of the filters. Each PLL provides an estimation of the instantaneous

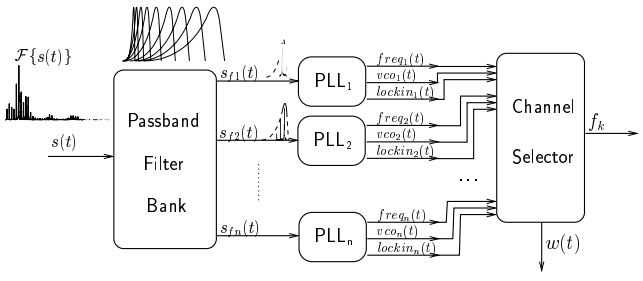

Fig. 3. Frequency channel splitter and selector diagram

phase ($vco_k(t)$) and frequency ($freq_k(t)$) of the main harmonic in each filter output signal. These PLLs outputs and the corresponding $lockin_k(t)$ signals are driven to the channel selector block. This stage samples $lockin_k(t)$ signals at a fixed frame rate and choose the channel with the maximum lock indicator value as the most reliable channel at that time. As was previously mentioned, lockin signal constitutes an indication of the degree of adjustment between the PLL phase estimation and the actual phase of the input. The corresponding frequency estimation of this selected channel f_k (which is $freq_k(t)$ at the sampling time) is assumed to be a reliable measure of the frequency of one harmonic, i.e. it is equal to the fundamental frequency multiplied by an integer factor i_k.

3.3. Spectrum Frequency Estimation block

Estimation of the coarse fundamental frequency $\hat{f}_0$ is accomplished by taking a section of the input signal $s_w(t)$ and performing Discrete Fourier Transform (DFT) on it. It is known that the DFT of a periodic windowed signal will have peaks concentrated near the harmonics of the fundamental frequency. The larger the section of the signal, the shaper those peaks are. However, because of the non stationary nature of the speech signal, we cannot indefinitely increase the width of the window without risk of averaging different portions of the signal (and also averaging fundamental frequency and consequently loosing resolution). We choose a window width equal to twelve times the inverse of the frequency f_k estimated above. We found experimentally that inside the allowed range of frequencies (50-500Hz), reasonable sharpness of the peaks of the DFT are obtained with that value. Finally we band limit the DFT to the samples representing frequencies lower than $2000\,Hz$, and perform the inverse of the DFT of this band limited spectrum. The coarse estimated frequency is measured in the resulting signal as the inverse of the time interval between zero time and the occurrence of the first peak.

3.4. Frequency comparator block

Now, we have a precise estimation of a harmonic of the fundamental frequency, but we do not know the harmonic number. On the other side we have a coarse estimation of the fundamental frequency. Using both estimations the block named Frequency Comparator finds the number of harmonic choosing the integer r that minimizes the difference between multiples of $\hat{f}_0$ and the frequency of the selected channel:

$$i_k = r \text{ such that } \min_r(r\,\hat{f}_0 - f_k), r \in \mathbb{N}$$

Once we have the factor i_k, we find f_0 by simply dividing f_k and i_k.

$$f_0 = f_k / i_k$$

4. EXPERIMENTAL RESULTS

4.1. Data and description of experiments

Performance is evaluated using Keele pitch extraction reference database [8]. Pitch reference is provided from simultaneously recorded laryngograph trace. It is available at <ftp://ftp.cs.keele.ac.uk/pub/pitch/>. It consists on 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 get_f0 algorithm [9], a well known pitch extraction algorithm available as part of **Wavesurfer** toolkit (see <http://www.speech.kth.se/wavesurfer/>). Frame rate is set to 10 msec and range frequency estimation from 50 to 500Hz in both our system and Wavesurfer. Other parameters of Wavesurfer are set to defaults.

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). Otherwise in the rest of voiced frames 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 ([3], [10], [11]), we divided voiced frames into two sets based on pitch estimated from Wavesurfer: "clearly voiced frames", where reference voiced segments are truly detected by Wavesurfer, and "voiced-to-unvoiced frames", where reference voiced segments are wrongly detected as unvoiced.

4.2. Results

Table 1 shows the performance of our system under clean conditions. 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 (clearly voiced plus voiced-to-unvoiced frames). In the later case we considered as gross errors those voiced frames detected as unvoiced by Wavesurfer.

Table 2 shows performance of the system under noisy conditions. In this case results on total voiced frames only are presented. 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. As a consequence a distinction between GER in total voiced frames and voiced to unvoiced errors is meaningless. Accuracy is measured with the addition of white noise to the signal with a SNR from 30db to 0dB, in 10dB steps.

4.3. Summary and discussion

Table 1 shows that both GER and MAE errors are nearly equal in both systems in the zone corresponding to frames detected

	Wavesurfer voiced frames		
Estimator	GER %	MAE (Hz)	STD (Hz)
Wavesurfer	2.101	2.60	3.91
PLL	**2.033**	**2.66**	**3.31**
	Total reference voiced frames		
Estimator	GER %	MAE (Hz)	STD (Hz)
Wavesurfer	6.315	2.60	3.91
PLL	**2.565**	**2.87**	**3.73**

Table 1. *Results for clean speech*

	Wavesurfer		
Noise Level(dB)	GER %	MAE (Hz)	STD (Hz)
30	6.66	2.58	3.87
20	9.04	2.50	3.68
10	21.11	2.33	3.08
0	64.49	2.33	2.44
	PLL		
Noise Level(dB)	GER %	MAE (Hz)	STD (Hz)
30	2.59	2.88	3.81
20	2.73	2.88	3.80
10	3.17	2.91	3.80
0	6.02	3.24	4.06

Table 2. *Results for noisy speech*

as voiced by Wavesurfer. However, when we consider total voiced reference frames we can see that Wavesurfer significantly increases GER while our system has a much lower increase. This result can be predicted by the lockin property of PLLs on which our system is based. Whenever a non-voiced to voiced change is produced 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.

Table 2 shows that the PLL based system measure of GER largely outperforms Wavesurfer pitch estimator. This is also a consequence of the PLL behavior. As said, PLLs lockin property is highly immune to noisy conditions. As a consequence, if no significant loss of lock of the PLL bank is produced, the rest of the system blocks will not be affected by the presence of noise, and no significant degradation of the pitch estimation will occur. This fact is shown in both GER and MAE measures where no significant degradation is observed in the PLL case. It is important to note that while Wavesurfer measure of MAE also do not have significant degradation, the number of voiced frames which are truly detected is significantly reduced as SNR decreases.

PLLs devices have many attractive features which are specially suitable for speech signals analysis. We have developed a system based on such devices which uses some of these features, including the lockin property with instantaneous phase and frequency in harmonics of the signal and noise robustness. This work shows that a pitch extraction system based on PLLs devices is more robust than the state-of-the-art.

5. REFERENCES

[1] W. Hess, *Pitch determination of speech signals*, Springer-Verlag, 1983.

[2] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer and intonation teaching," in *Eurospeech 1993*, pp. 1003–1006.

[3] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *IEEE, ICASSP 2000*, June.

[4] F. M. Gardner, *Phaselock Techniques*, John Wiley and Sons, 1979.

[5] Claudio Estienne and Patricia Pelle, "A synchrony front-end using phase-locked-loop techniques," in *ICSLP 2000*, Beijing, China, vol. III, pp. 98–101.

[6] Patricia Alejandra Pelle and Matias Capelettto, "Pitch estimation using phase locked loops," in *EUROSPEECH 2003*.

[7] K. Wang, Eyal Yair, and A. Shamma, "Auditory analysis of espectro-temporal information in acoustic signals," *IEEE Engineering in medicine and biology*, pp. 186–194, March 1995.

[8] F. Plante, G. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Eurospeech 1995*, pp. 837–840.

[9] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT)*, Elsevier, 1995.

[10] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "Robust pich estimation at very low snr exploting time and frequency domain cues," in *IEEE, ICASSP 2005*.

[11] Fei Sha, J. Ashley Burgoyne, and Lawrence K. Saul, "Multiband statistical learning for f_0 estimations in speech," in *IEEE, ICASSP 2004*.