A ROBUST PITCH DETECTOR BASED ON TIME ENVELOPE AND INDIVIDUAL HARMONIC INFORMATION USING PHASE LOCKED LOOPS AND CONSENSUAL DECISIONS

Patricia Pelle and Claudio Estienne

Institute of Biomedical Engineering and Electronic Engineering Department
School of Engineering, University of Buenos Aires
ppelle,cestien@fi.uba.ar
Paseo Colon 850, (1063) Ciudad Autonoma de Buenos Aires, Argentina

ABSTRACT

A pitch detector system is proposed, based on estimation of time envelope information and individual harmonics frequency determination. The system takes advantage of the fact that both the time envelope of a speech signal and the frequency of its individual harmonics carry information about the pitch. A set of pitch hypotheses are calculated from time envelope signals extracted from different parts of the spectrum. These hypotheses are tested against the information obtained from individual harmonics from the lower part of the spectrum. Hypotheses which reliably match that information are preserved, and a consensual decision is taken between them to obtain the final pitch estimation. In addition, several by-products of the process can be retained in order to perform voiced/unvoiced detection. As every subsystem extracts information from arrangements of Phase Locked Loops the detector is very robust in noisy conditions, compared to the well known get_f0 algorithm.

Index Terms— Pitch estimation, consensual decisions, PLL frequency estimation, PLL spectrum estimation, PLL noise robustness.

1. INTRODUCTION

Pitch determination is a complex problem. Many difficulties arise in pitch estimation, including pitch-doubling, pitch-halving, performance degradation with noise, voiced-unvoiced decision and estimation at the beginning and ending of voiced segments. Several algorithms have been proposed in the past [1] [2] [3] [4] [5] that perform very well over clean speech. But nearly all fail in noisy environmental conditions, discouraging their use in many speech processing systems. Nevertheless pitch is important in several applications, such as distinguishing segmental categories in tonal languages, speech coding systems, and speech analysis-synthesis systems among others, so the lack of a robust pitch detector is a pending issue.

In the past we have already developed two approximations to robust pitch detection [6, 7], both of them making use of Phase Locked Loop (PLL) devices [8]. In [6] we explored appealing properties of PLLs, such as the ability to automatically track periodic signals and extracting their instantaneous phase, even under severe noise conditions. Based on those key features we developed a pitch detection system that was able to outperform current systems under noisy conditions. In [7] we added a block which has the ability to make a consensual decision between many guesses obtained from the main system. This latest approach retained the robustness characteristics of [6], but adding increasing precision.

In this work we propose to include temporal envelope information to the previous approaches. Roughly speaking, the system produces several versions of the temporal envelope by using increasing spectral information ranging from 500 up to 5000Hz, and from every envelope instance a guess for the pitch value is obtained. All these guesses are tested against a secondary pitch detector based on individual harmonics of the lower part of the spectrum. Guesses that pass the test are retained to make a consensual decision about the true pitch value. This approach also produces a lot of collateral information related to the voiced/unvoiced characteristic of the input signal, that can be combined in a Linear Discriminant Analysis (LDA) discriminator providing voiced/unvoiced detection. Since the collateral information is derived from the main pitch detection, this voiced/unvoiced detector exhibits the same robustness characteristics than the pitch determination. This ability is also a key improvement over previous systems [6, 7].

2. OVERALL SYSTEM DESCRIPTION

In this section an overall system description is presented, leaving further detailed explanation of each part to the fol-

This work is funded by University of Buenos Aires grant: UBACYT 20020100100883.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

lowing section. The general block diagram is as shown in Figure 1. Three main stages compose the system. The block

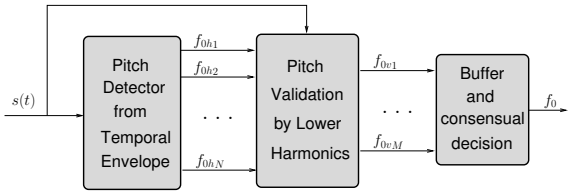

Fig. 1. General system block diagram

called Pitch Detector from Temporal Envelope computes several versions of the temporal envelope, and from each envelope instance estimates one value for the pitch f_{0h_n}. These estimations can be thought as initial hypotheses or guesses about the true pitch value. In the block termed Pitch Validation by Lower Harmonics, a test for each postulated pitch value is performed, scoring them by using information taken from the lower harmonics of the input signal. Only those pitch estimations that have an score exceeding a predefined threshold are retained and passed towards the following block, while other non successful estimations are discarded. The latest block accumulates validated pitch estimates, f_{0v_m}, for three consecutive frames, to perform a kind of consensual decision between them by calculating their median value. 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 emphasized that this approach of generating a lot of redundant values for the pitch and taking a consensual decision between the most highly qualified ones, produces a great reduction in the error rate (see [7]).

3. DETAILED DESCRIPTION

3.1. Pitch Detector from Temporal Envelope

The block that calculates pitch estimates from the temporal envelope is composed of several parallel blocks, all of them displaying the same structure shown in Figure 2. The main

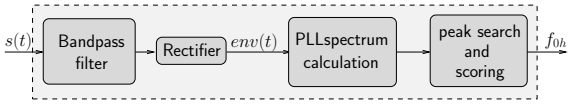

Fig. 2. Block diagram of pitch detection from temporal envelope

goal in the envelope computation is to obtain a signal that exhibits strong energy in the fundamental harmonic, stronger than the fundamental harmonic energy in the input signal itself. This objective can be attained by making a temporal rectification of the signal, as long as more than one harmonic of the signal is present. It can be consider approximately that rectification reconstructs the fundamental frequency of a signal. In order to obtain different envelope instances in each structure, the first component in the block is a bandpass filter whose high stop frequency is different for each of these structures in the whole block but wide enough to assure that is capturing more than one signal harmonic.

Once the rectification is performed, the rest of the block is devoted to obtain the frequency of this main harmonic component of the envelope signal $env(t)$. The method used to find this frequency is to perform a PLLspectrum calculation, described below, and to determine which one of the spectrum peaks more likely match the harmonic structure in this spectrum, by calculating a score for each peak as:

$$score = \sum_{k=0}^{K} PLLspectrum(k)\cos(2\pi k/p) \qquad (1)$$

where p is the discrete index corresponding to a peak in the spectrum. If this score is high, more harmonics of p are present, that indicates that the p is a good pitch candidate. So the output of this block is the frequency corresponding to the peak with the highest calculated score.

The PLLspectrum for a signal is a kind of spectrum computation using Phase Locked Loops [8]. The reason to use this particular spectral calculation is that it inherits the same noise robustness as PLLs themselves, as it was shown for example when it was used in the recognition of voiced sounds [9]. In this work a simpler version of that PLLspectrum is used, as it is depicted in Fig. 3. The first stage is a bank of bandpass

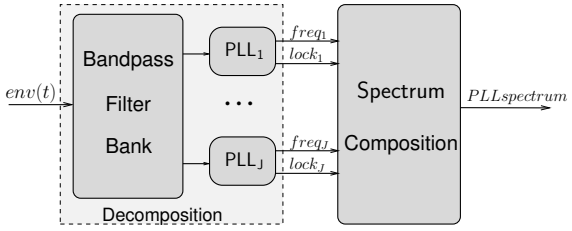

Fig. 3. Block diagram for PLLspectrum calculation

filters that decomposes the input signal into simpler signals, each filter followed by one PLL. The outputs of each PLL are the inputs to the Spectrum Composition stage. The bandpass filter at the beginning of the block is intended to restrict the frequency band to which PLLs should be able to synchronize. Band boundaries are linearly distributed in mel scale. The degree of asymmetry and Q was experimentally set, using as guide the biological descriptions of [10]. If the goal is capturing individual harmonics of the signal as in this case, narrow band filters are required. In the Spectrum Composition stage the PLLs outputs (frequency measured and lockin, see [8] [6] [7]) are averaged over a frame, and these averaged PLLs outputs are applied to generate a vector that describes energy as function of frequency. The objective is to register a measure

of which frequencies are mainly present in the PLL bank output, representing the frequency composition in the input signal. The set of these vectors can be arranged in the form of a pseudo-spectrum, similar to the power spectrum of the signal. The lockin output is used as a kind of energy weight for the PLL output frequency, because PLLs always displays a frequency output, even though there is not a sinusoidal signal in its input. The spectral composition consists on locating the lockin value of each PLL in the abscissa indicated by its frequency. If more than one PLL in the bank indicates the same frequency, their lockin values are added. As a final step, a linear convolutions with a smoothing window is performed over this vector, that also produces energy average for very similar frequencies.

The PLLspectrum calculated for this block is composed of 23 bandpass-filter/PLL pairs, with their frequency band upper limit ranging from 80 to 600Hz. Parameters for PLLs are experimentally set, as defined in previous works [9]. We implemented 200 parallel structures as shown in Figure 2, i.e. there are 200 f_{0h_n} initial hypotheses for the pitch value. The high stopband frequency for the bandpass filter in each block ranges from 500 to 5000Hz. All the bandpass filters are cochlear filters, as used by Wang and Shamma in [11].

Additional useful information for voice/unvoiced detection can be obtained from this block. Three parameters are extracted from each parallel block: the mean and variance of the PLLspectrum for each frame, and the score of the spectrum peak chosen as the pitch value f_{0h} of the block. Those parameters display a strong variation between voiced and unvoiced segments, that will be combined later with other parameters extracted from the validation block.

3.2. Validation by Lower Harmonic Information

This block mainly consist of recomputing the score as expressed in Eq. 1, but in this case another PLLspectrum is used, calculated directly from the lower spectral part of the input signal. This evaluation produces a kind of cross spectral check between different portions of the signal spectrum. As was previously mentioned, the time envelope of the signal is calculated from high portions of the signal spectrum. In contrast, in this block only low harmonics of the input signal are used. The comparison of this score against a predefined threshold is intended for reassurance about the the reliability of initial hypothesis calculated from the temporal envelope, selecting better quality f_{0h} values for the final consensual computation. This PLLspectrum ranges from 0 to 1000Hz, where it is possible to identify easily individual harmonics. Forty channels are used, with bandpass cochlear filters of the same class as before, PLLs values and other parameters established as in [7], and the reliability threshold experimentally set. Three additional parameters are retained for further use in voiced/unvoiced detection: the new score for the pitch candidate, and the mean and variance of the lower harmonics PLLspectrum.

3.3. Voiced/Unvoiced detection

The voiced/unvoiced detector is a simple Linear Discriminant Analysis (LDA) detector [12]. From every f_{0h_n} computation (including those discarded in the Validation Block), the collection of the six parameters previously described that will hopefully characterize the quality of sonority of the frame are extracted. Another additional parameter is added to the set, which is the ratio between the original f_{0h_n} and the final pitch obtained after the consensual decision between validated hypothesis. This latest parameter is generally equal to one in voiced segments, and very spread in unvoiced ones. The LDA is trained to give an answer for each f_{0h_n} about their quality as voiced. Also in this case a consensual decision is taken to obtain the final voiced/unvoiced decision, by assuming that the majority of answers for each f_{0h_n} voicing quality is the true value for voiced/unvoiced. In order to avoid overfitting, the LDA detector is trained with 4/5 of the signals in the data base and tested with the remaining portion. This process is repeated five times in order to cover all the signals in the database. For noisy signals, the whole set of clean signals was used to train the LDA detector.

4. EXPERIMENTS AND RESULTS

4.1. Experiments and data description

Performance is evaluated using two freely available databases, with simultaneously recorded laryngograph trace that provides a reference considered as the ground truth. The first database is called Keele pitch extraction reference database [13][1], that consists on five male and five female speakers, each speaking a short story of about 35 seconds. The second database was produced by P. Bagshaw [3][2], which is composed of fifty English sentences, each spoken by one male and one female speaker for a total of approximately 7 minutes each. Both databases are studio quality, sampled at 20 KHz. Noise added to evaluate the system for noisy conditions is taken from NOISEX database examples[3]. Results are compared to those of get_f0 algorithm [2], a well known pitch extraction algorithm included as part of **Wavesurfer** toolkit. Frame rate is set to 10 ms and range frequency estimation from 50 to 500Hz in both our system and Wavesurfer. Other parameters of get_f0 are set to their 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 this case). Two kind of gross errors are produced in a pitch determination system. On one hand,

[1] <ftp://ftp.cs.keele.ac.uk/pub/pitch/>
[2] <http://www.cstr.ed.ac.uk/research/projects/fda/>
[3] <http://spib.rice.edu/spib/selct_noise.html>

some voiced frames can be wrongly detected as unvoiced, and on the other hand, where frames are correctly detected as voiced, the pitch value can be outside the limits of non gross error. For some systems as get_f0 for example, it is not possible to disable the voiced/unvoiced detection, so these two error sources are jointly present in the measured GER. For our system we present two GER measures: gross errors over voiced frame as appears in the reference (GERtot), and gross errors produced by wrongly pitch determination plus wrongly voiced frames taken as unvoiced (GERvoiced). In the first GER measure, only one source of error is reported (wrong pitch values), whereas in the second the combination of both souces are reported, and in the case of get_f0 GER. Also the percentage of voiced frames detected as unvoiced (V2U) and the percentage of unvoiced frames taken as voiced (U2V) are reported for both systems, completing the description.

4.2. Results and discussion

Tables 1 and 2 show the GER performance of our system compared with get_f0, in clean and noisy conditions. Noise added is white, ranging from 30 to 0dB of SNR. GER comparison has to be done over the last two columns in both tables, in order to compare similar errors counts as was explained in the previous section. In both clean or low noise conditions, the two detections are similar (PLL system is slightly better for Bagshaw database, while get_f0 performs better for Keele database). But while the gross errors frames for the PLL system in severe noise is less than 10% more of the total frames of the low noise condition, get_f0 increases its errors by a percentage of more than 50. In Tables 3 and 4 voiced/unvoiced errors are shown for both systems. These two tables display the same kind of behavior as the previous case. While both PLL system and get_f0 present similar performance for clean and low noise (again, PLL system performs better for Bagshaw data base and get_f0 for Keele), the behavior for high noise is completely different. In the case of the PLL detector under severe noise conditions, the increase in the percentage of frames wrongly detected unvoiced (V2U) from clean conditions is only 1.5% for the Bagshaw database, and around 5% for Keele, while the performance of wrongly detected voiced frames (U2V) is maintained at around 3% of increase. But in the case of get_f0, the error count ocurring into the voiced part of input signal (V2U) is very high, misdetecting more than 50% of the frames voiced as unvoiced. PLL system is taking advantage of two factors: the good performance of the voiced/unvoiced detector, and the small percentage of total voiced frames with a gross pitch error (first column of Tables 1 and 2), resulting in a significant improvement over get_f0 algorithm.

SNR (dB)	GERtot PLL	GERvoiced PLL	GER get_f0
clean	2.27	4.96	5.71
30	2.35	5.05	5.79
20	2.34	5.38	6.81
10	3.14	7.01	14.55
0	7.67	10.7	56.67

Table 1. *Gross Error Rate for Bagshaw database*

SNR (dB)	GERtot PLL	GERvoiced PLL	GER get_f0
clean	3.18	7.28	6.28
30	3.22	7.56	6.63
20	3.34	8.32	9.01
10	4.55	10.96	21.10
0	10.01	16.08	64.49

Table 2. *Gross Error Rate for Keele database*

SNR (dB)	V2U PLL	U2V PLL	V2U get_f0	U2V get_f0
clean	3.31	8.35	3.81	5.23
30	3.34	8.29	3.99	3.90
20	3.71	8.02	5.17	2.41
10	4.80	7.89	13.53	0.78
0	4.89	11.84	56.48	0.05

Table 3. *Voiced/unvoiced errors for Bagshaw database*

SNR (dB)	V2U PLL	U2V PLL	V2U get_f0	U2V get_f0
clean	4.97	8.60	4.30	7.38
30	5.23	8.39	4.72	5.93
20	5.93	8.12	7.48	3.73
10	8.20	7.74	20.15	1.26
0	10.12	9.00	64.44	0.17

Table 4. *Voiced/unvoiced errors for Keele database*

5. CONCLUDING REMARKS

The performance of the PLL pitch detector presented in this work is similar to that of a good conventional algorithm as get_f0 for clean signals, while largely outperforms its behavior for noisy conditions, especially for severe noise environment. The key points in this behavior are the intensive use of PLLs and the redundancy of the system. This redundancy is not only due to the fact that multiple values are calculated each time, but also to the fact that the values are composing information provided by different parts of the spectra. While the proposed system is both more time and memory consuming than get_f0, higher performance largely justify this extra complexity.

6. REFERENCES

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

[2] D. Talkin, *Speech Coding and Synthesis*, chapter A Robust Algorithm for Pitch Tracking (RAPT), pp. 495–518, Elsevier Science Inc., 1995.

[3] 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*, 1993, pp. 1003–1006.

[4] Chao Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP'00.*, 2000, vol. 3, pp. 1343–1346 vol.3.

[5] Alain de Cheveign and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[6] P.A. Pelle, "A robust pitch extraction system based on phase locked loops," in *International Conference on Acoustics, Speech and Signal Processing, ICASSP'06.*, Toulouse, France, May 2006, vol. 1, pp. I–I.

[7] Patricia Alejandra Pelle and Claudio Francisco Estienne, "A pitch extraction system based on phase locked loops and consensus decision," in *International Conference on Speech communication and technology (INTERSPEECH 2007)*, Antwerp, Blgica, Ago 27-31 2007, ISSN 1990-9772.

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

[9] P. Pelle, C. Estienne, and H. Franco, "Robust speech representation of voiced sounds based on synchrony determination with PLLs," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2011. (ICASSP 2011) Proceedings.*, Prague, Czech Republic, May 22–27 2011, pp. 5424–5427.

[10] M. I. Miller and M. B. Sachs, "Representation of stop consonants in the discharge patterns of auditory-nerve fibers," *The Journal of the Acoustical Society of America*, vol. 74, no. 2, pp. 502–517, 1983.

[11] K. Wang and S.A. Shamma, "Auditory analysis of spectro-temporal information in acoustic signals," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 14, no. 2, pp. 186–194, Mar/Apr 1995.

[12] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, Springer, corrected edition, August 2003.

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