Robust Pitch Tracking of Acoustic Signals based on Local Periodicity Index

Roudra Chakraborty
Jadavpur University, Kolkata.
roudra@gmail.com

Debapriya Sengupta
Indian Statistical Institute, Kolkata.
dps@isical.ac.in

Abstract

In this paper, a method of pitch tracking based on variance minimization of locally periodic subsamples of an acoustic signal is presented. Replicates along the length of the periodically sampled data of the signal vector are taken and locally averaged sample variances are minimized to estimate the fundamental frequency. Using this method, pitch tracking of any text independent voiced signal is possible for different speakers over any database.

1. INTRODUCTION

Extraction or determination of fundamental frequency (or pitch period) of a speech signal is a fundamental problem in both speech processing and speaker recognition. The typical pitch range for a male human being is 80-200 Hz, and for females 150-350 Hz. Many methods to extract the pitch of speech signals have been proposed. Improvement in accuracy of performance, robustness against noise of these methods are still desired. Unfortunately, we do not have very reliable and accurate method for pitch extraction in noisy environments. Also measuring the period of a speech waveform, varying in and with the detailed structure of the waveform, can be quite difficult. Another problem is automatic selection of the window of the voiced speech segments.

Correlation based methods [10] and [17] are known to be comparatively robust against noise. The autocorrelation function method and the average magnitude difference function (AMDF) method [16] are classified into this category.

Based on these methods there are some other techniques like auditory modeling [3], probabilistic AMDF modeling [5], fourier transform modeling [8], real-time digital hardware pitch detector [13], semiautomatic pitch detector (SAPD) [14], automatic formant analysis [15], modified autocorrelation and AMDF [18], projection measure technique [19], pseudo-pitch synchronous analysis [20] and many more [12]. Ideas on pitch extraction have also been discussed in some tutorials [4] and [7].

In this paper, we propose a new method for extraction of fundamental frequency of speech signal using simple statistical techniques. Statistical characteristics of sample variances calculated over periodic subsamples have been used to extract fundamental frequency from speech signals. Similar ideas for pitch tracking can be found in [2], [6], [9] and [11].

The remainder of this paper is organized as follows. Section 2 describes the principle of the proposed method. In Section 3, we show the results of preliminary test for the proposed method and compare with some standard pitch detection methods we confirm the effectiveness of our method. Then finally in Section 4, we conclude this paper giving views regarding further development that can be done.

2. PROPOSED METHOD

Consider a voiced segment $y = (y_1, y_2, \ldots, y_n)$ in a digital speech signal. Since most speech signals can be viewed as a quasi-periodic sequence the fundamental frequency may not be uniquely defined mathematically. In our approach we estimate the fundamental frequency by statistically enhancing the most significant harmonics present in y.

For a given time period k consider a periodically subsampled data

$$y_{b,w} = \{y_b, y_{b+1}, y_{b+2}, \ldots, y_{b+w}\},$$

where b is an index varying from 1 to $w(\left[\frac{n}{w}\right] - 1)$ and w is taken to be a constant.

Let us define the following parameters as

$$m = \left[\frac{w+1}{2}\right] - 1,$$

$$S_{i,k} = \left\{i, i+k, \ldots, i+pk : p \in Z^{+}\right\},$$

where

$$k = 1, 2, \ldots, m;$$

$$i = 1, 2, \ldots, k;$$

p is the greatest integer such that, $i + pk \leq b + w$.

Now we define our local periodicity index (LPI) as

$$g(k) = \frac{1}{k}\sum_{i=1}^{k}\left[\frac{1}{p+1}\sum_{j\in S_{i,k}}\left\{y_{b,w}(j) - \overline{y}_{b,w}(j)\right\}^2\right], \tag{1}$$

where

$$\overline{y}_{b,w}(j) = \frac{1}{p+1} \sum_{j \in S_{i,k}} y_{b,w}(j).$$

Here

$$\sigma_i^2 = \frac{1}{p+1} \sum_{j \in S_{i,k}} \left\{ y_{b,w}(j) - \overline{y}_{b,w}(j) \right\}^2$$

leads to a chi-square test if the signal is too noisy, otherwise it is normally distributed for clean signal.

Let f_0 be the sample rate of the original speech signal and

$$f(k) = \frac{f_0}{k}.$$

In view of (1) g can be thought of as a function of f. We want to find the largest significant minimizer. If f^* is the chosen minimizer, we can expand the signal in the harmonics $\{f^*, 2f^*, 3f^*, \ldots\}$. Larger values of f^* introduces parsimony in the description of the signal. In order to define an automatic criterion for identification of such an f^* we define the following function $c(f)$ as,

$$c(f) = \sum_{j=1}^{m} \left\{ \frac{f^*(j)}{f} - \left[\frac{f^*(j)}{f} \right] \right\} exp\left\{-Ag(j)\right\} \qquad (2)$$

Here in $c(f)$ decimal part of $\frac{f^*(j)}{f}$ has been weighted with the exponential function of g, so that when g takes larger values or $\frac{f^*(j)}{f}$ takes near-integral values, $c(f)$ decreases rapidly. $c(f)$ is a vector of length k (A being a positive constant). Let i be the index of the minimum of the components of $c(f)$. Then $f^*(i)$ is referred as the fundamental frequency of the speech signal of y.

3. EXPERIMENTS

A. Experimental Details

Clean and noisy speech signals were obtained from IViE Corpus [1]. Speech samples were uttered by two male and two female speakers. Each of such speech signals consisted of a maximum of five English words, which were sampled by different rate.

B. Preliminary Test

Taking $w = 1000$ and $A = 50$, we found data sets of f^* and the following graphs :

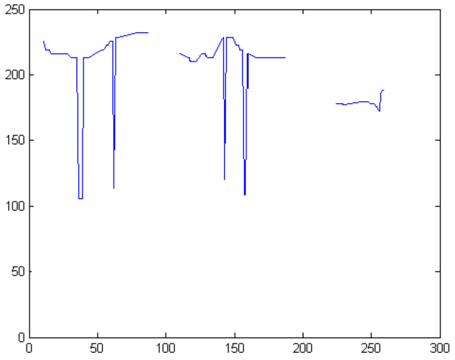

Figure 1 : Graph for fundamental frequency of speaker F1

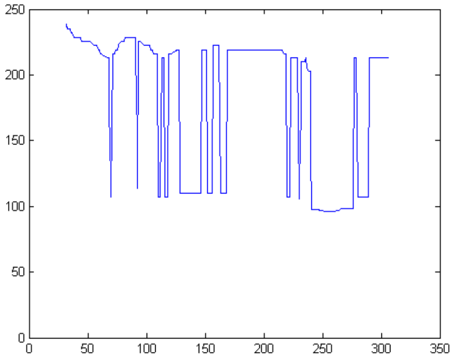

Figure 2 : Graph for fundamental frequency of speaker F2

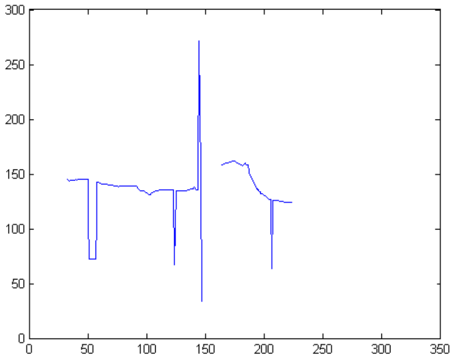

Figure 3 : Graph for fundamental frequency of speaker M1

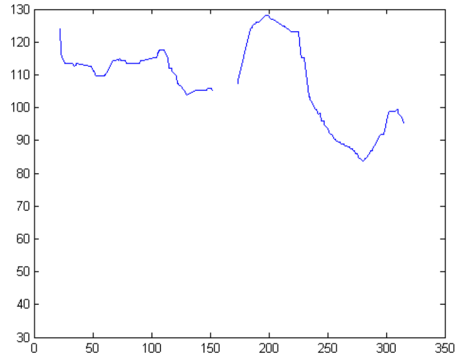

Figure 4 : Graph for fundamental frequency of speaker M2

C. Comparison with Some Standard Methods

To investigate the accuracy of the LPI, we have conducted experiments which compare it with two conventional method. The conventional methods are the methods of Autocorrelation and MFCC.

Autocorrelation method does the correlation analysis frame-by-frame to the estimated average pitch period of the speaker. Given a discrete time signal $x(n)$, the autocorrelation function is defined as

$$r_x(m) = \frac{1}{n}\sum_{i=0}^{n-1} x(i)x(i+m),$$

defined for all n and lag m. The property of this function is that $r_x(m)$ is large when $x(n)$ has similar value with $x(n+m)$. If $x(n)$ has a pitch period ρ, then $r_x(m)$ has peaks at the integral multiples of ρ. Obviously $r_x(0)$ is maximum among these values, the second largest being $r_x(\rho)$. Other maximals usually decrease as m increases. Therefore using this method we can estimate ρ from the location of the peak at $m = \rho$.

MFCC uses the full-band speech signal for processing. MFCC designs a mel-scaled filterbank for the supplied number of channels and processes the input speech signal through the filterbank into a set of mel-scaled cepstral coefficients. It uses a Hamming window and a forward FFT and and a specialised inverse DFT.The window size determines the lowest frequency one can still capture with a cepstrum. The formula for the minimal frequency is: $2 \times (sampling\ rate/window\ size)$.

If a signal was recorded with a sampling rate of 22050 Hz, a 1024 point window can cope with fundamental frequencies down to 43 Hz, while an analysis with 128 points will not work for frequencies below 344 Hz. Pitch detection will generally improve with larger windows.

In [12] a detailed comparative performance study has been discussed of these two methods.

The readings suggest that the proposed method is competitive with the other two methods. Although in practical situations we face noisy speech in most of the cases, robustness of LPI doesn't take much care of the fact whether the environment is noisy or not. The graphs are as follows:

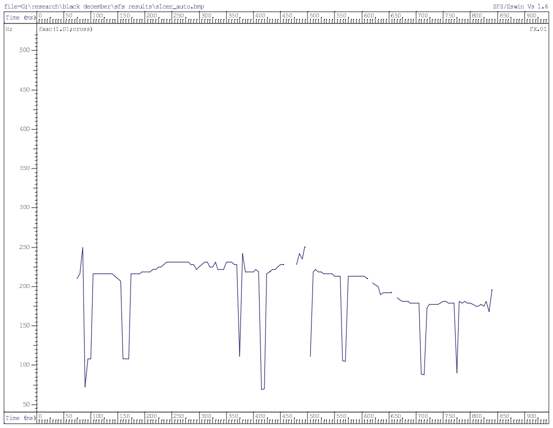

Figure 5 : Graph for fundamental frequency of speaker F1 using Autocorrelation

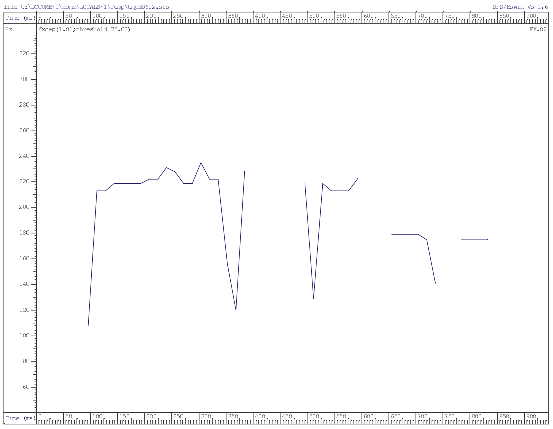

Figure 6 : Graph for fundamental frequency of speaker F1 using MFCC

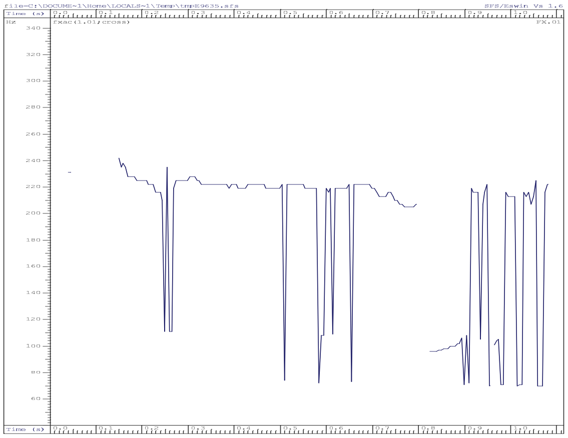

Figure 7 : Graph for fundamental frequency of speaker F2 using Autocorrelation

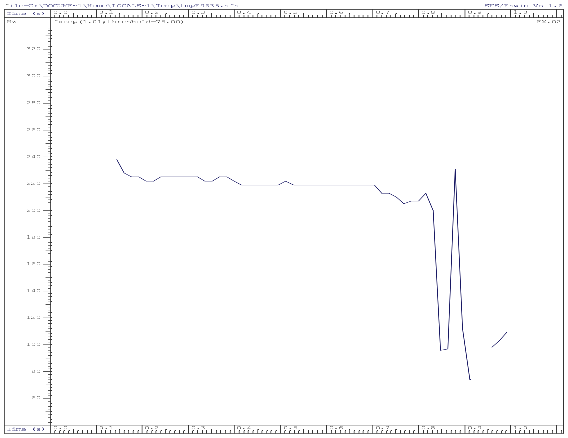

Figure 8 : Graph for fundamental frequency of speaker F2 using MFCC

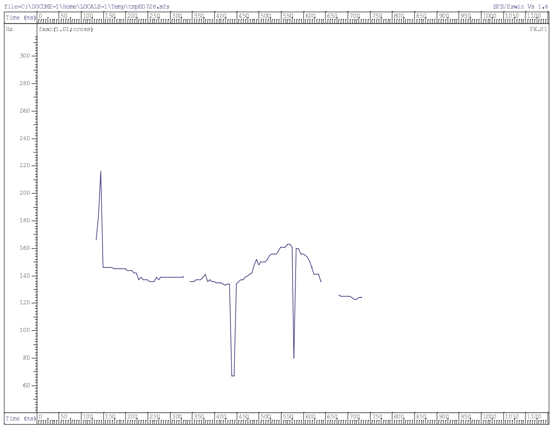

Figure 9 : Graph for fundamental frequency of speaker M1 using Autocorrelation

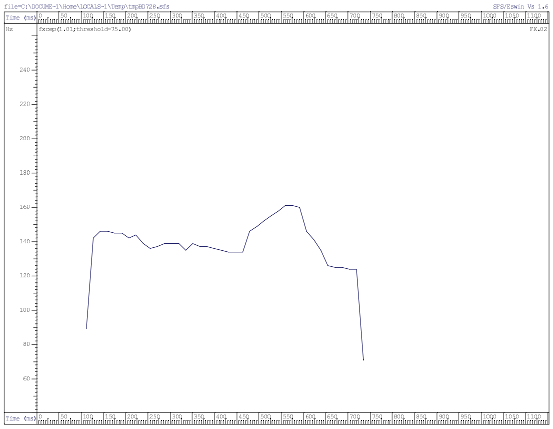

Figure 10 : Graph for fundamental frequency of speaker M1 using MFCC

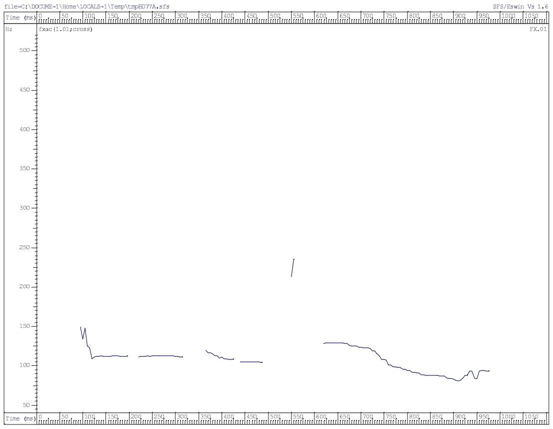

Figure 11 : Graph for fundamental frequency of speaker M2 using Autocorrelation

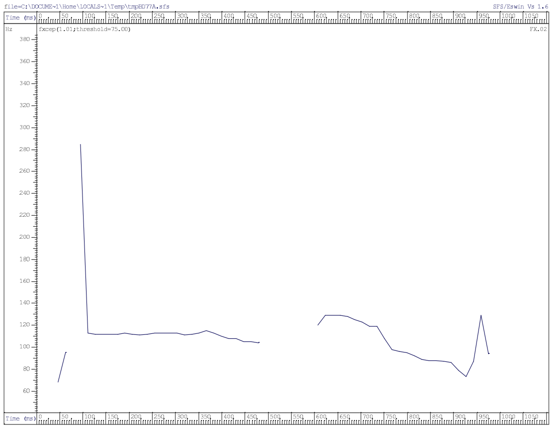

Figure 12 : Graph for fundamental frequency of speaker M2 using MFCC

4. CONCLUSION

Based on the experimental results it has been shown that LPI is useful in both noisy and clean environment. The idea behind this method also leads to methods of extraction of other features of speech signals. The limitation of this method is the procedure to obtain the minimum values from the data set which are often used to fall behind the known limits viz. 80-200 Hz for males and 150-350 Hz for females.

REFERENCES

[1] *The IViE Corpus.* Phonetics Laboratory, University of Oxford and Department of Linguistics, University of Cambridge, 1997-2002.

[2] D. G. Childers. *Speech Processing and Synthesis Toolboxes.* John Wiley & Sons Inc. 2000.

[3] P. Cosi, S. Pasquin and E. Zovato. *Auditory Modeling Techniques for Robust Pitch Extraction and Noise Reduction.* Proc. of ICSLP (1998) paper 1053, 1998, citeseer.ist.psu.edu/cosi98auditory.html.

[4] David Gerhard. *Pitch Extraction and Fundamental Frequency: History and Current Techniques.* Technical Report TR-CS 2003-06:1-22, November 2003.

[5] Leah H. Jamieson, Goangshiuan S. Ying and Carl D. Michell. *A Probabilistic Approach to AMDF Pitch Detection.* Proceedings of the 1996 International Conference on

Spoken Language Processing, Philadelphia, PA:1201-1204, October 1996.

[6] F. Jelinek. *Statistical Methods for Speech Recognition.* The MIT Press, Cambridge, MA, 1998.

[7] Joseph P. Campbell Jr. *Speaker Recognition: A Tutorial.* Proceedings of the IEEE, 85(9):1437-1462, September 1997.

[8] Sylvain Marchand. *An Efficient Pitch-tracking Algorithm Using a Combination of Fourier Transforms.* Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland:DAFX1-DAFX6, December 2001.

[9] T. F. Quatieri. *Discrete-Time Speech Signal Processing.* Pearson Education Pte. Ltd, 2004.

[10] L. R. Rabiner. *On the use of Autocorrelation Analysis for Pitch Detection.* IEEE Trans. Acoust., Speech, Signal Processing, ASSP-25(1):24-33, February 1977.

[11] L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proc. IEEE, 77(2):257-286, February 1989.

[12] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal. *A Comparative Performance Study of Several Pitch Detection Algorithms.* IEEE Trans. Acoust., Speech, Signal Processing, ASSP-24(5):399-417, October 1976.

[13] L. R. Rabiner, J. J. Dubnowski and R. W. Schafer. *Real-Time Digital Hardware Pitch Detector.* IEEE Trans. on Acoustics, Speech, and Signal Processing, ASSP-24(1):2-8, February 1976.

[14] L. R. Rabiner, C. A. McGonegal and A. E. Rosenberg. *A Semiautomatic Pitch Detector (SAPD).* IEEE Trans. on Acoustics, Speech, and Signal Processing, ASSP-23(6):570-574, December 1974.

[15] L. R. Rabiner and R. W. Schafer. *System for Automatic Formant Analysis of Voiced Speech.* Journal of the Acoustical Society of America, 47(2):634-648, February 1970.

[16] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley. *Average Magnitude Difference Function Pitch Extractor.* IEEE Trans. Acoust., Speech, Signal Processing, ASSP-22:353-362, October 1974.

[17] T. Shimamura and H. Kobayashi. *Weighted Autocorrelation for Pitch Extraction of Noisy Speech.* IEEE Trans. on Speech and Audio Processing, 9(7):727-730, October 2001.

[18] L. Tan and M. Karnjanadecha. *Pitch Detection Algorithm: Autocorrelation Method and AMDF.* Proceedings of the 3rd International Symposium on Communications and Information Technology, 2:551-556, September 2003.

[19] Kuo-Hwei Yuo, Tai-Hwei Hwang and Hsiao-Chuan Wang. *Comparison of Autocorrelation-based Features and Projection Measure Technique for Speaker Identification.* IEEE Trans. on Speech and Audio Processing, 13(4): 565-574, July 2005.

[20] R. D. Zilca, B. Kingsbury, J. Navratil and G. N. Ramaswamy. *Pseudo Pitch Synchronous Analysis of Speech with Applications to Speaker Recognition.* IEEE Trans. on Speech and Audio Processing, DOI 10.1109-TSA.2005.857809, 2006.