Optimal Pitch Path Tracking for more reliable Pitch Detection*

Petr Motlíček and Jan Černocký

Brno Univ. of Technology, Inst. of Radioelectronics,
{motlicek,cernocky}@urel.fee.vutbr.cz

Abstract. This paper presents an application of one method for improving fundamental frequency detection from the speech. The method is based on searching the best pitch paths over one or more words. It uses the idea that the fundamental frequency of a speaker cannot change sharply in a short time so that the pitch should not vary rapidly over one (or a few) words. This technique is created for improving the pitch detection. It cannot detect the pitch itself, but it uses some pitch detectors. We compare some of them here and we try to determine which one is the most suitable for our method.

1 Introduction

The word pitch (in context of speech processing), as defined operationally by psychoacousticians, is the frequency of pure tone that is matched by the listener to a more complex (usually periodic) signal. This is a subjective definition. When we speak of a "pitch detector", we usually refer to a device that measures the fundamental frequency of an incoming signal; this is an objective definition. Pitch detection and fundamental frequency estimation are often used interchangeably. The pitch detection is not such an easy task. We have a lot of difficulties, which we can encounter [1]:

- Sub-harmonics of fundamental frequency often appear that are sub-multiples of the "true" frequency.
- Fundamental frequency changes with time, often with each glottal period.
- In many cases when strong sub-harmonics are present, the most reasonable objective pitch estimate is clearly at odds with the auditory percept.
- It is difficult to distinguish periodic background noise from breathy voiced speech.
- The dynamic range of the voice fundamental frequency is large. The pitch of some male voices can be as low as 50 Hz, whereas the pitch of children's voice can be as high as 600 Hz.

* This work is supported by the Ministry of Education, Youth and Sports of the Czech Republic – project No. VS97060, and by the Research programme of Brno University of Technology "Research of electronic communication systems and technologies", No. CEZ: J22/98:262200011.

These different kinds of behavior make pitch detection difficult.

In the literature, a lot of techniques for pitch detection can be found. Among essential we can count autocorrelation function (ACF) [4]. Instead of ACF, normalized cross correlation function (NCCF) can be used [2]. NCCF is often able to estimate the pitch better than an ordinary ACF. Another method uses linear prediction LPC and the fundamental frequency is estimated from its error signal. In these techniques the pitch can be found using long-term prediction (LTP). Cepstral pitch detection is popular, as well. Here the pitch is estimated from the cepstral train [3], [5].

Using one of pitch detectors mentioned above only, we almost never get good pitch estimation. We find out that these techniques often have problems with recognition of voiced and unvoiced parts of speech. Moreover they often detect wrong values of pitch in speech (some higher harmonics or sub-harmonics), even though the fundamental frequency is still approximately unvarying. Therefore some other techniques are applied in order to remove these drawbacks. One of often used methods is the oversampling of input signal (usually 6 times). Another one is the median filtering of obtained pitch sequence. Finally, one of the ways that can improve pitch estimation is the optimal pitch path detection.

2 Pitch detector experiments

All methods that can work as pitch detectors were used in the experiments. We determined their ability to detect the pitch on about twenty Czech words (two speakers). On the whole it can be said, that the best pitch estimation is obtained using NCCF. It gives the biggest probability that we find out the "true" pitch. One example is given on Figure 1. NCCF, ACF, cepstrum and ACF of LPC-error signal are shown for one frame of input speech signal. The correct value of lag for this frame is around 90 samples. As it is shown in Figure 1, the best estimation of "true" pitch is obtained using NCCF, because of the biggest peak with respect to other peaks of its function. The other detectors failed to estimate the "true" pitch (caused by sub-harmonics of input signal).

2.1 Techniques for improving pitch estimation

As mentioned before, we are not able to achieve good pitch estimation using uniquely pitch detectors. Therefore some other techniques are used that can improve this estimation. These techniques are applied at the sequence of lags (one lag refers to one frame of speech signal). First possibility is the median filtering of lags' sequence. By this filter we can remove some wrong estimation of pitch in frames. It can also repair some wrong determination of one or a few voiced or unvoiced frames in lags' sequence. Some problems can be caused by a small sampling frequency of speech signal. So an oversampling of the input signal (mostly 6 times) can bring better pitch estimation.

We have found out that estimating of voiced or unvoiced frame is not so problematic. Problems are mostly caused by sub-harmonics of the "true" fundamental frequency in signal that are obtained after using some pitch detectors.

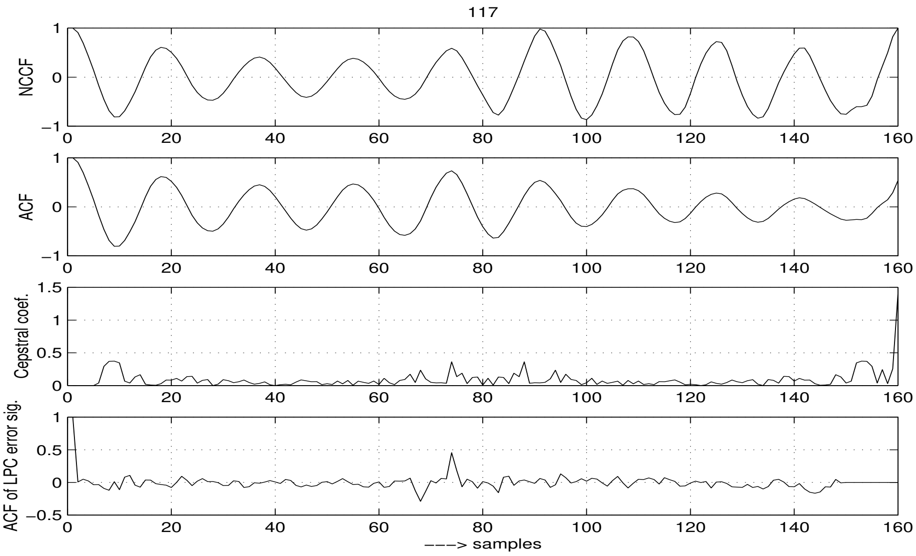

Fig. 1. Pitch detectors and their output functions of 117^{th} frame of input speech signal (two Czech words "létající prase".)

These drawbacks cannot be removed by previous methods. When ACF is computed (or other mentioned functions) we often find out that the interesting interval contains two, three, or sometimes more peaks, caused by sub-harmonics of the pitch. So if only one maximum of ACF is looked for, some sub-harmonics of fundamental frequency could be determined instead of the "true" pitch, because it can have bigger value than the real fundamental frequency. These drawbacks can be eliminated by using method which looks for optimal pitch path in lags' sequence.

2.2 Optimal pitch path detector

This detector looks for the suitable paths of lags in the input signal of n frames. First of all, a matrix $\mathbf{A}$ is given. This matrix contains a parameterization (ACF, NCCF, ...) for each frame in one row (so that it has n rows). In each row, 4 peaks are found out and their positions are recorded in matrix $\mathbf{M}$ (n times 4). The corresponding values are stored in matrix $\mathbf{V}$ (n times 4). Next, the unvoiced frames are detected: each element of $\mathbf{M}$ is taken and we determine if the following or the previous row of $\mathbf{M}$ contains a similar value (in any column). So if for i-th frame any of $m_{i,k}$ satisfies:

$$m_{i,k} \in m_{i-1,m} \pm const, \qquad \forall i,k, \qquad m \in [1,4] \tag{1}$$

or

$$m_{i,k} \in m_{i+1,m} \pm const, \qquad \forall i,k, \qquad m \in [1,4], \tag{2}$$

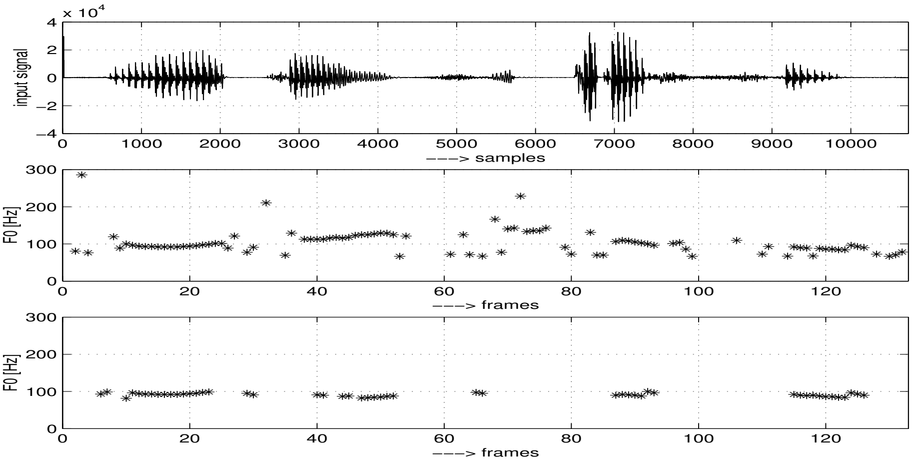

Fig. 2. Estimation of fundamental frequency of two Czech words "létající prase".
Upper panel: Input signal (10781 samples, 8kHz, 133 frames).
Middle panel: Pitch estimation using only NCCF.
Lower panel : Estimation using optimal pitch path technique (applied on NCCF).

where *const* is a constant, the i-th frame is declared voiced, otherwise it is unvoiced.

The next step is seeking all paths in *voiced portions* of matrix $\mathbf{M}$. A path is determined by its mean lag value stored in vector $\mathbf{b}$. This vector is initialized to empty. When processing new row of $\mathbf{M}$, its entries are compared to elements of $\mathbf{b}$. If an entry $m_{i,k}$ lies closed enough to an element b_i, it is clustered to b_i and b_i is updated using:

$$b_{i(new)} = \frac{b_{i(old)} + m_{i,k}}{2}. \tag{3}$$

If not, a new entry in $\mathbf{b}$ is created: $\mathbf{b} = [\mathbf{b}\ m_{i,k}]$. It is obvious that the length of $\mathbf{b}$ is not a-priori known. A vector $\mathbf{c}$ associated to $\mathbf{b}$ contains a sum of respective maxima values from the matrix $\mathbf{V}$. This vector therefore contains a "quality measurement" of pitch paths. At the end of a voiced portion, values of $\mathbf{c}$ are normalized by path lengths and the optimal mean lag b_d is determined:

$$d = \arg\max_i(\mathbf{c}), \qquad i \in [1, m], \tag{4}$$

where m is the number of candidate paths. The optimal path associated with b_d is extracted from $\mathbf{M}$. Finally, a simple energy detector marks the "unvoiced" frames where the energy falls below a pre-determined threshold.

The estimation of fundamental frequency of speech using optimal pitch path detection in comparison with raw NCCF, can be seen on Figure 2. This method is likely to recognize voiced and unvoiced parts of speech successfully.

3 Evaluation of results

Optimal pitch path detection was tested on a reference database from OGIvox[1] (Center for Spoken Language Understanding) that include simultaneous speech and electroglottograph (EGG) data. The OGIvox database contains a directory with signal files and a directory with "pitch ticks" (obtained by EGG) associated to these signal files. We have created a script that processes one thousand of signal and EGG files from OGIvox.

First, the real pitch is computed from the "ticks" for each frame of signal files, whereas the pitch is determined as a time period between the two closest "ticks". However it could happen that a frame contains only one, or more than two "pitch ticks". Therefore, we compute the pitch using "ticks" not only from the current frame, but also by using two "ticks" from two neighboring frames. Then time periods are determined and the mean value represents the pitch of this frame.

Second, the pitch was computed by using optimal pitch path detection for all files and results were statistically compared:
The error of detection of voiced frames in comparison with OGIvox (1.000 signal files) is 22.5%.
The error of detection of pitch in voiced frames is 9.6%.
Comparison of results for one OGIvox signal file is shown on Figure 3.

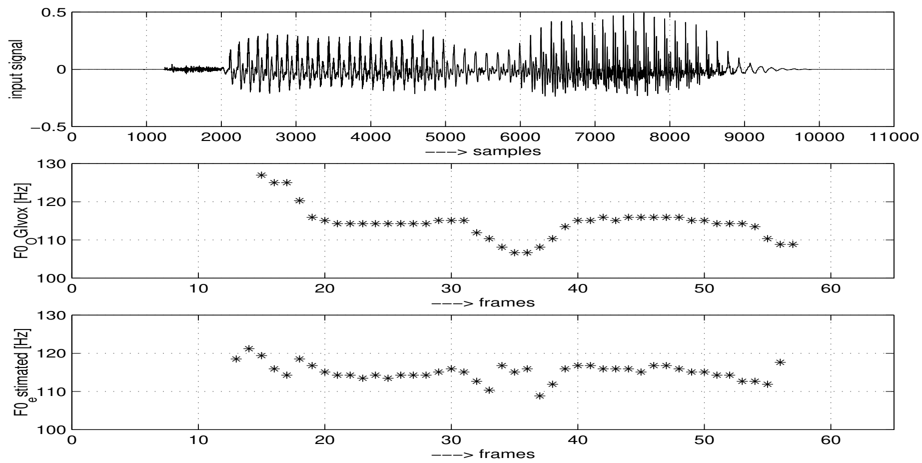

Fig. 3. Estimation of fundamental frequency F0 on signal file from OGIvox database:
Upper panel: Input signal (10880 samples, 16kHz, 67 frames).
Middle panel: "Real" pitch computed from EGG.
Lower panel : Estimation using optimal pitch path technique (applied on NCCF).

[1] Thanks to Mike Bacon (OGI, Portland, Oregon) for the permission to use it.

4 Conclusions

We have tested several methods of improvement the pitch detection. In general the optimal pitch path detection gives the best results. The biggest problems in pitch detection are caused by higher harmonics or sub-harmonics of the fundamental frequency. Only optimal pitch path detection of all mentioned techniques is able to find out the "true" fundamental frequency in speech signal. This method works with short sections of speech signal (a few words), thus it is able to recognize the change of pitch of a speaker not immediately, but after processing of one section of speech signal. The disadvantage of this technique is its higher complexity in comparison with other methods. The best pitch detector, which can be used with the optimal pitch path technique, is based on NCCF.

Optimal Pitch Path Detection was tested on Ogivox database (pitch provided by EGG). Determining of "true"pitch in voiced frames was very good (error 9.6% for computing over 1.000 signal files of speech). In detection of voicing, the error was 22.5%. However, a very simple detector of energy was used in the test, a more sophistical detector should improve this figure.

References

1. B. Gold and N. Morgan. *Speech and Audio Signal Processing*, pages 214-227 and 415-430, New York, 1999.
2. D. Talkin. A Robust Algorithm for Pitch Tracking (RAPT). In Kleijn, W. B. and Paliwal, K. K. (Eds.), *Speech Coding and Synthesis.* New York: Elseviever, 1995.
3. B. P. Bogart, M. J. R. Healy, and J. W. Tukey, The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and shape tracking, in *Symphosium on Time Series Analysis* (M. Rosenblatt, ed.), (New York), pp. 209-243, John Wiley and Sons, 1963.
4. L. R. Rabiner, On the use of autocorrelation analysis for pitch detection, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp.24-33, February 1977
5. J. Cernocky. *Speech Processing Using Automatically Derived Segmental Units*, PhD Thesis, ESIEE, France, 1998.