ROBUST PITCH DETECTION ALGORITHMS FOR ESTIOMATION OF FUNDAMENTAL FREQUENCY OF PROLONGED VOWELS PHONATIONS AT PATHOLOGICAL VOICES

L. Bauer[1], J. Rusz[1, 2], R. Čmejla[1]

[1]Czech Technical University in Prague – FEE, Department of Circuit Theory, Czech Republic

[2]Charles University in Prague, First Faculty of Medicine, Department of Neurology and Centre of Clinical Neuroscience, Czech Republic

Abstract

This paper presents a design for two novel methods of speech fundamental frequency (f_0) detection in pathological voices. The methods are tested on the database of vowel-sustained phonation. One method is based on the detection of maxima and the other on signal filtration with band pass. Compared to other well known f_0 detection methods, our algorithms are designed with respect to speech pathology detection. Moreover, the designed methods support detection of the other voice parameters, such as jitter, shimmer and harmonic-to-noise ratio (HNR). The comparison of our results with results generated by the help of the Praat algorithm reaches a high value of correspondence: for the maximum method, the accuracy is 88,4 % and for the pass band method it is 83,9 %. The detection leads to the creation of a self-automated method, which is robust in detecting f_0.

1 Introduction

Calculation of the fundamental frequency (f_0), which consists of vocal cord vibration, is an important step in many speech applications. The speech signal $x(t)$ does not consists exclusively of the fundamental frequency $s(t)$, but it is influenced by noise $n(t)$ which rises in the vocal cord, or in the background where is it recorded (see Equation 1). Consequently, the speech signal in a large scale is not purely periodic, but there is a change of irregular periods. Among patients with disordered speech, this noise (aperiodic) component of the signal increases and the detection of f_0 becomes a complex task.

$$x(t) = s(t) + n(t) \tag{1}$$

The fundamental frequency detection algorithms have extensive application in actual practice: for example, when creating synthetic voices, in speaker voice recognition, in linguistics to detect prosody, and equally in medicine, as a simple and fast approach to detect speech pathology and to control the progress of its treatment [1].

Commonly used f_0 detection algorithms are not usable for detection of speech disorders, because they do not identify the required moments in the speech signal. With the finding of the moments in the speech signal, we can determine other vocal parameters that are then used to determine the voice pathology. Ranked among these parameters are jitter (frequency instability), shimmer (amplitude instability) and HNR (harmonic-to-noise ration) [2]. For the detection of these parameters, it is necessary to have each signal period detected precisely and accurately. Existing speech pathology detection algorithms are not yet sufficiently precise or self automated.

For this reason, it is necessary to propose a new robust and self automated f_0 detection algorithm, which will lead to accurate detection of other speech parameters by which it will be possible to detect speech pathology.

2   The fundamental frequency detection algorithms

In general practice, two f_0 detection approaches are dominant. The first approach is called the peak picking method [3], which means pitch detection in the signal, which is represented by a commercial Multi-Dimensional Voice Program. The second method is called waveform matching [4], which tries to find the similarity between separate periods in the signal. The waveform matching method is represented by the freely accessible software Praat. In this paper we will detect f_0 by the use of the peak picking method.

The signal must first be preprocessed. Each signal preprocessing step is identical for both algorithms. It is important to note that the signal filtering affects the original signal, so it may smooth some important information for us. For this reason, we use only the DC component subtraction. The next step in the signal preprocessing is to find the beginning and the end of speech. The detection of the beginning and the end of the speech is based on the integration envelope of the signal. When it exceeds the defined threshold, the beginning and end of the speech is recognized. Due to the suppression of pitch doubling and pitch halving of the fundamental, the average f_0 throughout the signal is detected. It is determined from the Welch power spectral density, where we are looking for values between 60 and 500 Hz. In this range there is a common occurrence f_0 [5].

The first method of f_0 detection (see Fig. 1) is based on the detection of peaks (maxima) in the segmented signal (see step 1). The segment length is chosen as 1/20 of the average fundamental frequency once it is found. The segment size is selected with respect to the computational cost of the process, and for avoidance of overlapping of closely spaced peaks in the selected segment. In the detected maxima, only local maxima are kept. In some cases, there is a jump on the wrong pitch, so we try to avoid comparing the detected position of the pitch with the detection of a new peak between the two previously found peaks (see step 2, 3). At the beginning and the end of the speech, there is a start-up and finishing of the vocal cords. For this reason, we begin to detect the position of peaks from the center of the recording into the edges (see step 4), where the vocal cord vibration is already settled. Correctly detected peaks are the peaks with a higher sum of amplitudes (see step 5 ◐).
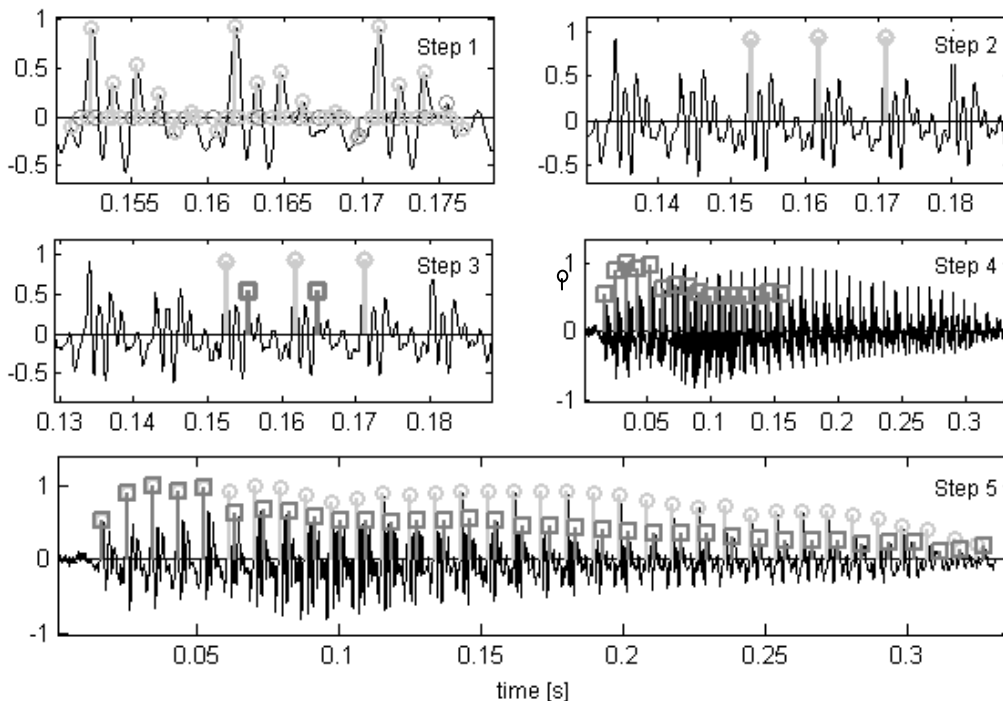

Figure 1: The maxima pitch detection method

The second f_0 detection method (see Fig. 2) is based on the segmentation of the signal and its subsequent filtration through band-pass 5th-order Butterworth filter. The length of the particular segment is 6000 samples with the 2000 samples overlap. In each segment the average f_0 value is again found, and around this frequency (± 10 Hz) the segment band pass is then filtered. Filtering here causes suppression of formant frequencies and noise. In the filtered signal the positions of peaks are then

detected and in their vicinity the positions of maxima in the signal are specified. This method is expected to involve a lower computational cost of the process, but the filtering of individual segments is relatively computationally demanding.
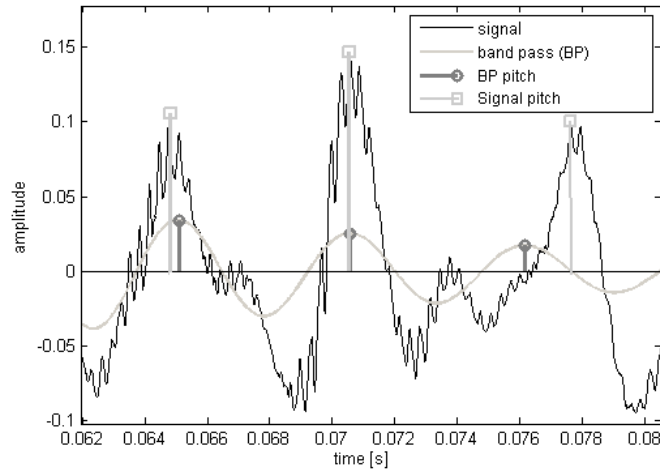

Figure 2: The band pass filtration pitch detection method

3   Database and experiments

The recording database was recorded at the First Faculty of Medicine using a condenser microphone, video camera Panasonic NV-GS 180. The microphone was placed at a distance of 15 cm from the mouth of the speaker. The signal is sampled at 48 kHz and for signal conversion a 16-bit converter was used. The recording consists of prolonged phonation of vowels /a/, /i/ and /u/. For comparison of function f_0 detection algorithms, abbreviated recordings were used, which were labeled in the correct position of the speech signal found by the world-renowned online-supplied program Praat [4]. The database contains 63 speech samples and 70 healthy subjects speaking as well as pathological specimens. Their length ranges up to 0.5 seconds. This length appears to be sufficient for testing, and can be expected to achieve the same or even better results for long signals. Comparing the results with the results of both algorithms detected in the Praat program is shown in Table 1.

Table 1: Comparison of results with the Praat

Method	Maxima	Band pass
Accuracy	88,4 [%]	83,9 [%]
SD	14,77 [%]	17,83 [%]

4   Evaluation

This paper presents two new self-automated fundamental frequency detection algorithms. These algorithms are also used to detect other vocal parameters. The newly designed algorithms work by the method known in the literature as peak-picking. The accuracy of algorithms is more than 83% in the labeled database.

Inefficiency in the algorithm approach to a signal problem lies in the ability to respond to the variability amplitudes of peaks due to the positions found in previous periods. Here we offer a solution for assessing the individual peak positions due to the distance from the estimated position of the pitch. To improve the accuracy of the algorithms, multiple symptoms should be detected in different periods, so that period positions are not determined only by peaks, but on a larger scale, for example, the instant of passing the signal through zero or the point of the vocal cords closing interval.

References

[1] W. Hess. *Pitch Determination of Speech Signals*: *Algorithms and Devices*. Springer-Verlag, Berlin, 1984.

[2] J.I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco and F. Cruz-Roldan. *The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders*. Journal of Voice, vol. 24, number 1, pages 47-56, 2010.

[3] Kay Elemetrics Corp.: Multi-Dimensional Voice Program (MDVP) Model 5106: Software Instruction Manual. Lincoln Park, Kay Elemetrics, 2003.

[4] P. Boersma and D. Weenink. *Praat: doing phonetics by computer (Version 5.1.17)* [Computer program] Retrieved October 5, 2009, from http://www.praat.org/.

[5] J. Uhlíř, P. Sovka, P. Polák, V. Hanžl, R. Čmejla. *Technologie hlasových komunikací*. Nakladatelství ČVUT, Praha, 2007.

Lukáš Bauer
Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, Czech Republic
e-mail: bauerlu3@fel.cvut.cz

Jan Rusz
Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, Czech Republic

Charles University in Prague, First Faculty of Medicine, Department of Neurology and Centre of Clinical Neuroscience, Czech Republic
e-mail: ruszjan@fel.cvut.cz

Roman Čmejla
Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, Czech Republic
e-mail: cmejla@fel.cvut.cz