A Comparative Analysis of Pitch Detection Methods Under the Influence of Different Noise Conditions

Lyudmila Sukhostat and Yadigar Imamverdiyev, *Baku, Azerbaijan*

Summary: Objectives/Hypothesis. Pitch is one of the most important components in various speech processing systems. The aim of this study was to evaluate different pitch detection methods in terms of various noise conditions.
Study Design. Prospective study.
Methods. For evaluation of pitch detection algorithms, time-domain, frequency-domain, and hybrid methods were considered by using Keele and CSTR speech databases. Each of them has its own advantages and disadvantages.
Results. Experiments have shown that BaNa method achieves the highest pitch detection accuracy.
Conclusions. The development of methods for pitch detection, which are robust to additive noise at different signal-to-noise ratio, is an important field of research with many opportunities for enhancement the modern methods.
Key Words: Pitch detection–Time-domain methods–Frequency–Domain methods–Hybrid methods.

INTRODUCTION

Pitch originates due to vocal fold vibration, and the frequency at which the vocal folds vibrate is the fundamental frequency. It is an important attribute of voiced speech. Periodicity related to voiced speech segments is determined as “pitch step” in the time domain and as “fundamental frequency” or F_0 in the frequency domain.

Sounds from different sources are combinations of different frequencies. The lowest frequency of a harmonic complex sound is called the fundamental frequency.

In addition to providing valuable information about the nature of the excitation source for speech production, pitch contour is used for speaker identification, emotion recognition, voice activity detection tasks, and speech training for hearing-impaired people and is needed for almost all speech synthesis systems.[1]

The reason for the difference in speakers pitch is the size, mass, and tension of the vocal folds. The child’s fundamental frequency averages about 250 Hz, and the length of the vocal folds is about 10.4 mm. With age, the length of male vocal folds grows to approximately 15–20 mm, and female voice folds to 13–15 mm. These changes in size are correlated with a decrease in fundamental frequency. Women have a higher frequency range than men as they have a smaller larynx size.

However, accurate and reliable measurements of the pitch period from only one acoustic signal are often extremely difficult. In some cases, vocal tract formants can significantly change the structure of laryngeal signal so that the actual fundamental frequency is difficult to detect (during rapid movements of the articulators when the formants are also changing rapidly[2]). Another problem can occur in a noisy speech environment. In such cases, the detailed structure of the signal can be corrupted, which leads to an incorrect measurement of fundamental frequency. It can be very hard to distinguish between speech and nonspeech (silence, noise, music, or a variety of other acoustical signals such as door knocking, coughing, paper shuffling, and so forth). In more complex environments (street, shopping mall, café, and so forth), it can be very difficult to detect the speech signal of interest because of high noise level. Especially, if the background contains babble noise, whose statistical characteristics are very similar to speech, such as in an exhibition hall.

As a result of multiple difficulties of pitch measurement, different methods of its detection have been developed and compared Rabiner et al[2] and Veprek et al.[3]

Mainly, a pitch detector is a device, which makes voiced/unvoiced decisions. It ensures measurement of fundamental frequency in voiced segments; although in the majority of algorithms such as decision making, it is a part of the measurement process, not a detached stage.

In recent years, alternatives to more traditional approaches to pitch detection have been proposed.[4–7]

In this article, we provide an overview of existing approaches for the determination of fundamental frequency to identify their strengths and weaknesses.

This article describes time-domain, frequency-domain, and hybrid methods for pitch detection. Evaluation measures of methods are presented, and a brief description of speech databases is provided. The results and discussion of reviewed methods are also presented.

PITCH DETECTION METHODS

Existing pitch detection methods[2] can be divided into three groups:

- The first primarily uses the time-domain characteristics.
- The second uses the properties of frequency domain.
- The hybrid methods combine properties of both time- and frequency-domain characteristics.

Time-domain methods are performed directly on the speech signal.[8–10] For this type of pitch detectors[11–16] peaks, valleys, frequency of zero-crossings, and autocorrelation measurements are very important. Time-domain measurements will provide a

Accepted for publication September 12, 2014.
This work was supported by the Science Development Foundation under the President of the Republic of Azerbaijan – Grant number EİF-RİTN-MQM-2/İKT-2-2013-7(13)-29/18/1.
From the Department of Information Security, Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan.
Address correspondence and reprint requests to Lyudmila Sukhostat, Institute of Information Technology, Azerbaijan National Academy of Sciences, 9, B. Vahabzade Street, Baku AZ1141, Azerbaijan. E-mail: lsuhostat@hotmail.com
Journal of Voice, Vol. 29, No. 4, pp. 410-417
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.09.016

good evaluation of fundamental frequency if a quasiperiodic signal is appropriately processed to minimize the effects of the formant structure.

The methods from this category consist in consideration of the input signal as amplitude fluctuations in the time domain and detection of repetitive segments in the sound wave, which determines its periodicity.[17,18]

The following pitch detectors can be selected in the time domain: parallel processing time-domain method[13,19] and data reduction method,[11,14] where the error rate does not depend on the level of noise. These two methods have a slightly lower resolution than other methods due to the sensitivity to sound wave peaks, valleys, and zero-crossings to changes in formants, noise, distortions, and so forth. Addition of a smoothing algorithm improves their performance.

Difficulties of time-domain methods for the male voice with low-pitch tone occur due to the analysis of fixed length frames at 30–40 milliseconds which, as a rule, is not suitable for fundamental frequency evaluation.

Frequency-domain pitch detectors[20–23] use the property that if the signal is periodic in the time domain, then the frequency spectrum of the signal will consist of a series of impulses at the fundamental frequency and its harmonics.[2]

A typical analysis of frequency spectra consists in segmentation of the speech signal into small frames, their multiplication by a window function, and obtaining short-term Fourier transforms for each frame. If the signal is periodic, the Fourier transform shows several peaks corresponding to fundamental frequency.

Because human perception[1] is mostly logarithmic, this means that low pitches can be detected less accurately than high.

Such methods include Harmonic Product Spectrum (HPS),[24] cepstrum pitch determination,[25] and Linear Predictive Coding (LPC).[26] Advantages of the HPS method are that it is less computationally expensive and noise robust.[27] LPC[28] uses a feedback filter to separate signal excitation from the vocal tract. This method considers the real cepstrum of the speech signal for pitch detection.

A hybrid pitch detector is used in both the time and frequency domains. For example, hybrid detectors can use frequency-domain methods to provide temporary spectral aligned sound waves and then uses autocorrelation measurements to estimate the pitch period. This category analyzes the output of the band-pass filters in the time domain. In many cases, the algorithm identifies many pitch candidates for each frame and then uses time limitations to find the pitch contour.

Among the various pitch detection algorithms, the following methods can be selected: modified autocorrelation method (AUTOC),[2] cepstrum method, multiband summary correlogram (MBSC),[6] BaNa,[7] YIN,[4] yet another algorithm for pitch tracking (YAAPT),[5] average magnitude difference function (AMDF),[2] SWIPE,[29] and pitch estimation filter with amplitude compression (PEFAC).[30]

The following section describes these pitch estimation methods in more detail.

METHODS IN TIME DOMAIN

Modified autocorrelation method

The autocorrelation approach is the most widely used time-domain method for signal pitch period evaluation.[1] This method[15] is based on identification of the highest values of the autocorrelation function in the areas of our interest. A short-time autocorrelation function for a signal $s(n)$ $\{n = 0, 1, \ldots, N-1\}$ is calculated as following:

$$R(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} s(n)s(n+k), \quad k = 0, 1, \ldots, N-1, \tag{1}$$

where N is the frame length and k is the lag index.

According to (Eq. 1), we search for the peak, location of which determines the value of the pitch.

For female voices, cross-correlation and autocorrelation methods give similar results.[22] Nonetheless, both methods are not sufficiently robust, especially for very low-frequency speech.

AUTOC algorithm is as following:

Input: vector $x \in R^{1xN}$; speech sampling rate Fs; frame length L_m; frame shift R_m.

Output: extracted pitch period scores.

1. Low-pass filtering at 900 Hz
2. Converting L_m and R_m

$$L = L_m \cdot Fs/1000, \quad R = R_m \cdot Fs/1000$$

3. Find first one-third samples $I_{pk}1$ and last one-third samples $I_{pk}2$ for each frame
4. Compute the clipping level $C_L = \alpha \cdot \min(I_{pk}1, I_{pk}2)$
5. Calculate modified autocorrelation function (Eq. 1)
6. Find maximum autocorrelation peak at zeroth interval
7. Voiced/unvoiced decision
8. Median filtering

The main limitation of pitch evaluation using the autocorrelation method is the presence of autocorrelation peaks that exceed the peaks corresponding to the pitch period. As a result, we get "picking" of peaks, and consecutively incorrect pitch evaluation can occur.

Average magnitude difference function

The pitch detection algorithm using the AMDF[11] method has a relatively low computational cost and is easy to implement. The principle of speech signal pitch detection by using AMDF is based on a short-term function, which calculates the difference between the initial function of the speech signal and the time-shifted function.

AMDF is used in speech signals for pitch period measurement of voiced signals which are semiperiodical (pitch period is not defined for unvoiced signals because they do not have periodical excitation).

AMDF[6] function is calculated as following: