Method for estimating pitch independently from power spectrum envelope for speech and musical signal

Yoshifumi Hara[a*], Mitsuo Matsumoto[b] and Kazunori Miyoshi[a]

[a] Graduate school of engineering, Kogakuin University, Tokyo, Japan
[b] SV research associates, Fukuoka, Japan
*corresponding author: kuro5hiron@ymail.plala.or.jp

(Received 30 November 2009; accepted 7 December 2009)

Pitch and changes in pitch are primary characteristics of a speech signal. Since a speech signal is a quasi-periodic signal, stability and accuracy are required to a pitch estimation method. Various methods for extracting periods of a speech signal in the time domain and for analyzing the microstructure of the spectrum in the frequency domain have been proposed. Auto-correlation function (ACF) and its applications are well known methods to be applied to detect periodicity of a speech signal in the time domain and are known to be robust to noise. ACF in the time domain is equivalent to the power spectrum in the frequency domain. Therefore, pitch estimated by ACF is subject to the power spectrum of the speech signal.
This paper proposes a method for applying ACF to detect periodicity of the microstructure of the spectrum in the frequency domain, independently from the power spectrum envelope. First, divide a speech signal into a set of frames. Second, in each frame, picking up major local peaks of the amplitude frequency characteristics for a speech signal in the frame in the frequency domain. Third, represent the amplitude frequency characteristics as a sequence of unity impulses, which is a line spectrum. Locations of the impulses on the frequency axis are those of the local peaks. Finally, apply ACF to the sequence for extracting periods of the impulses on the frequency axis. And estimate pitch with the periods. Since pitch estimated by this method is free from the power spectrum envelope of a speech signal, the method has stability and accuracy. Furthermore, in this method, because simplified ACF is applicable to a line spectrum, the method is advantageous for computing complexity.

Key words: Pitch estimation, Autocorrelation, Power spectrum, Line spectrum, Peak-picking

1. INTRODUCTION

Pitch estimation for speech signal has long history. Pitch is a primary feature of a speech signal and a variety of methods for estimating pitch has been proposed. Definitive algorithm, however, has not been proposed. Pitch is unstable and a speech signal is subject to frequency characteristics of a vocal tract. Those are the difficulties in pitch estimation. For example in the time domain, a method for detecting a time interval between events, for example zero crossing which is a feature of a waveform, is sensitive to instability [1].

Auto-correlation Function (hereafter, ACF) is a well-known method for detecting periodicity of a signal. A speech signal is quasi- periodic and is generally composed of its fundamental frequency component and some overtones. Amplitude of the fundamental frequency component and those of the overtones are time-variant and the amplitude of the fundamental frequency component is not always dominant to those of the overtones. In some conditions, the fundamental frequency component is missing. Therefore, pitch of a speech signal estimated by ACF is subject to power spectrum of the signal.

On this issue, inverse filtering for power spectrum envelope, which is "spectrum whitening", is proposed [2][3]. Furthermore, some countermeasures in the time domain are proposed [4]. Power spectrum of a periodic signal has some peaks. The peaks correspond to its fundamental frequency and those of the overtones. Period estimation of the signal in the time domain is equivalent to extract intervals between the peaks on the frequency axis. ACF is applicable to extract the intervals. Even if ACF is applied in the frequency domain, the intervals detected by ACF are subject to powers of the peaks, because in almost all conditions, the peaks have different powers. Cepstrum is applied to extract the microstructure of the power spectrum independently from its power spectrum envelope. It's considered to be "spectrum whitening". For a musical signal, a probability density function is applied to extract the most predominant fundamental frequency supported by the harmonics [5].

In this paper, more intentional "power spectrum flattening" is developed by introducing unit-line spectrum representation of a musical sound. Here "Unit-line" denotes that every "line" has the same unit power, which means that the power spectrum envelope is "flat". Peak-picking, which is a method to extract dominant frequency component on the least-square error criterion in the time domain [6], is adopted to extract the "lines". Furthermore, auto-correlation analysis is applied to extract intervals on the frequency axis between the "lines". Results of fundamental frequency analysis for musical tones of a piano are compared with other conventional methods such as spectrogram according to short-term Fourier transforms, auto-correlation functions in the time domain, and envelope correlation analysis based on a pitch-sensation model [7].

2. FUNDAMENTAL FREQUENCY ANALYSIS FOR PIANO TONES

Four types of methods are compared in the experiments. Procedures for fundamental frequency analysis shown in Fig.1-

Fig. 4. Figure. 5 shows a sample of peak-picking. In this paper, Peak-picking is done six times in all the peak-spectrum analysis.

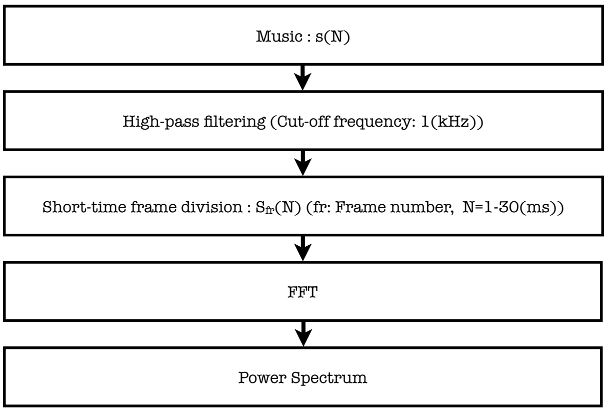

Music : s(N)
High-pass filtering (Cut-off frequency: 1(kHz))
Short-time frame division : $S_{fr}(N)$ (fr: Frame number, N=1-30(ms))
FFT
Power Spectrum

Fig. 1. Spectrogram

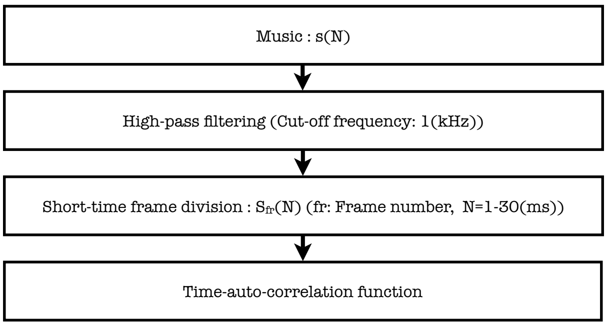

Music : s(N)
High-pass filtering (Cut-off frequency: 1(kHz))
Short-time frame division : $S_{fr}(N)$ (fr: Frame number, N=1-30(ms))
Time-auto-correlation function

Fig. 2. Time-auto-correlation function (ACF)

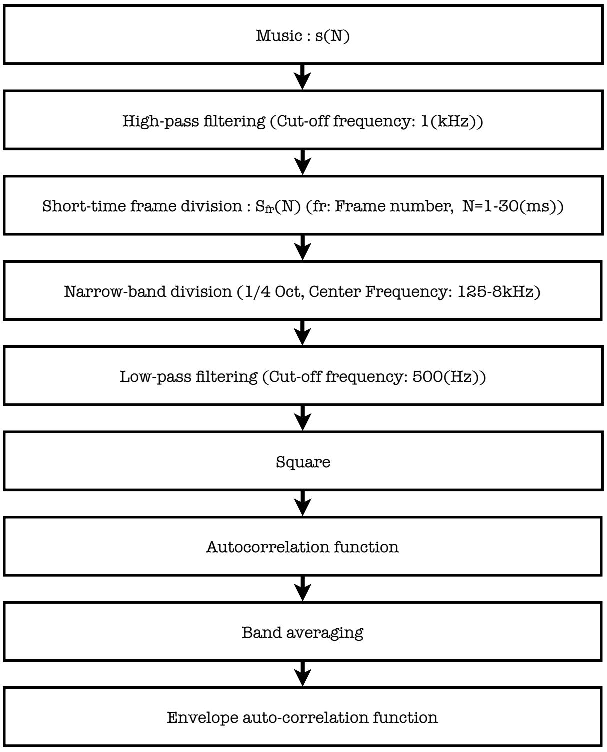

Music : s(N)
High-pass filtering (Cut-off frequency: 1(kHz))
Short-time frame division : $S_{fr}(N)$ (fr: Frame number, N=1-30(ms))
Narrow-band division (1/4 Oct, Center Frequency: 125-8kHz)
Low-pass filtering (Cut-off frequency: 500(Hz))
Square
Autocorrelation function
Band averaging
Envelope auto-correlation function

Fig. 3. Envelope ACF

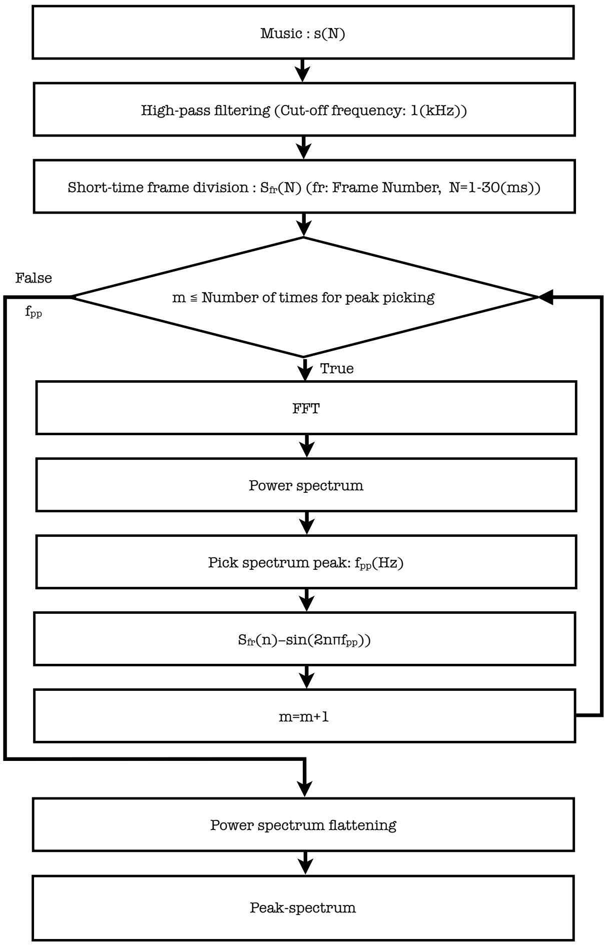

Music : s(N)
High-pass filtering (Cut-off frequency: 1(kHz))
Short-time frame division : $S_{fr}(N)$ (fr: Frame Number, N=1-30(ms))
False
f_{pp}
m ≦ Number of times for peak picking
True
FFT
Power spectrum
Pick spectrum peak: f_{pp}(Hz)
$S_{fr}(n)-\sin(2n\pi f_{pp})$
m=m+1
Power spectrum flattening
Peak-spectrum

Fig. 4. Peak-spectrum ACF

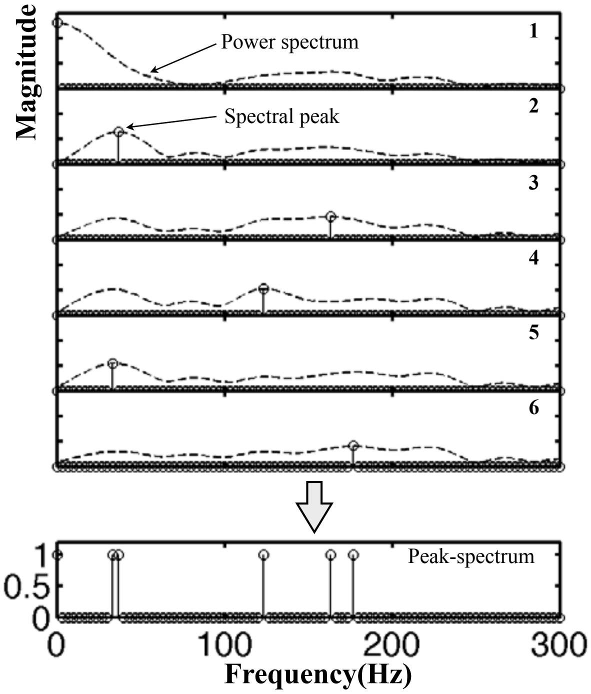


Fig. 5. A sample of spectral peak-picking

3. RESULTS

Figures 6 and 7 illustrate the results. Figure 6 presents the frequency analysis for a single tone of a piano (note A4≅440(Hz)).

Here high-pass type of filtering is applied(cut-off frequency: 1(kHz)) before analysis. This is because it is important to estimate the fundamental frequency even if there is no fundamental frequency, so-called under missing-fundamental conditions. Figure. 7 demonstrates the results for a compound sound like a chord (note A4≅440(Hz)+C# 5≅554(Hz)).

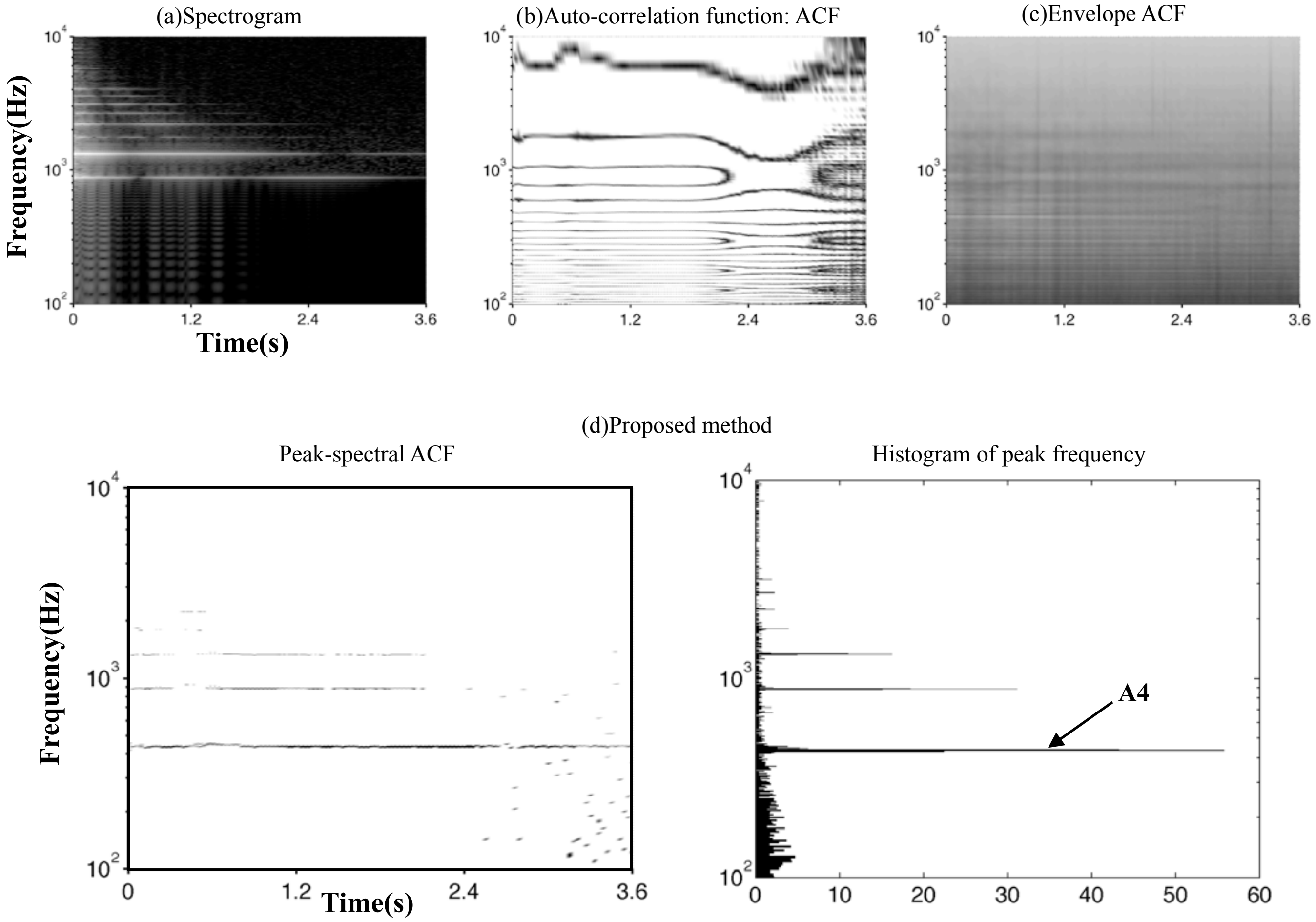
In both Figures panel(a) is the case for spectrogram, panel (b) shows the results by conventional ACF analysis, panel(c) presents the envelope correlation analysis, and panel(d) is obtained by the proposed method. Furthermore, right one of the panel (d) in Fig. 6 shows a histogram of peak frequency in left one of the panel(d). The fundamental frequency cannot be estimated by the spectrogram as shown in the panel(a). This is because such a frequency band which includes the fundamental frequency is contained in the signal to be analyzed. The fundamental frequency is not also clearly seen as shown in the panel(b).

On the other hand the envelope-correlation method which was developed inspired by a model of hearing organ may reveal that it might not be suitable for the fundamental frequency analysis on the short-term basis. Actually the method is well known as a good estimator of a fundamental frequency for a relatively long-stationary time interval rather than the short intervals in this example.

As far as we see these results, the fundamental frequency can be clearly estimated by the proposed method. This is because spectral "peak-picking" works well even for a very short time interval independent of spectral deformation effects of time windowing on the spectral analysis [6]. And moreover the proposed method is independent of strength of each spectral peak, but only the locations of spectral peaks on the frequency axis is utilized.

However it can be found that there are some difficulties in the case of a compound sound as shown in Fig. 7(d). There are some sub-harmonics in the result.

Such a sub harmonics implies that a further step would be necessary to a good and robust estimator of the fundamental frequency.



(a)Spectrogram

(b)Auto-correlation function: ACF

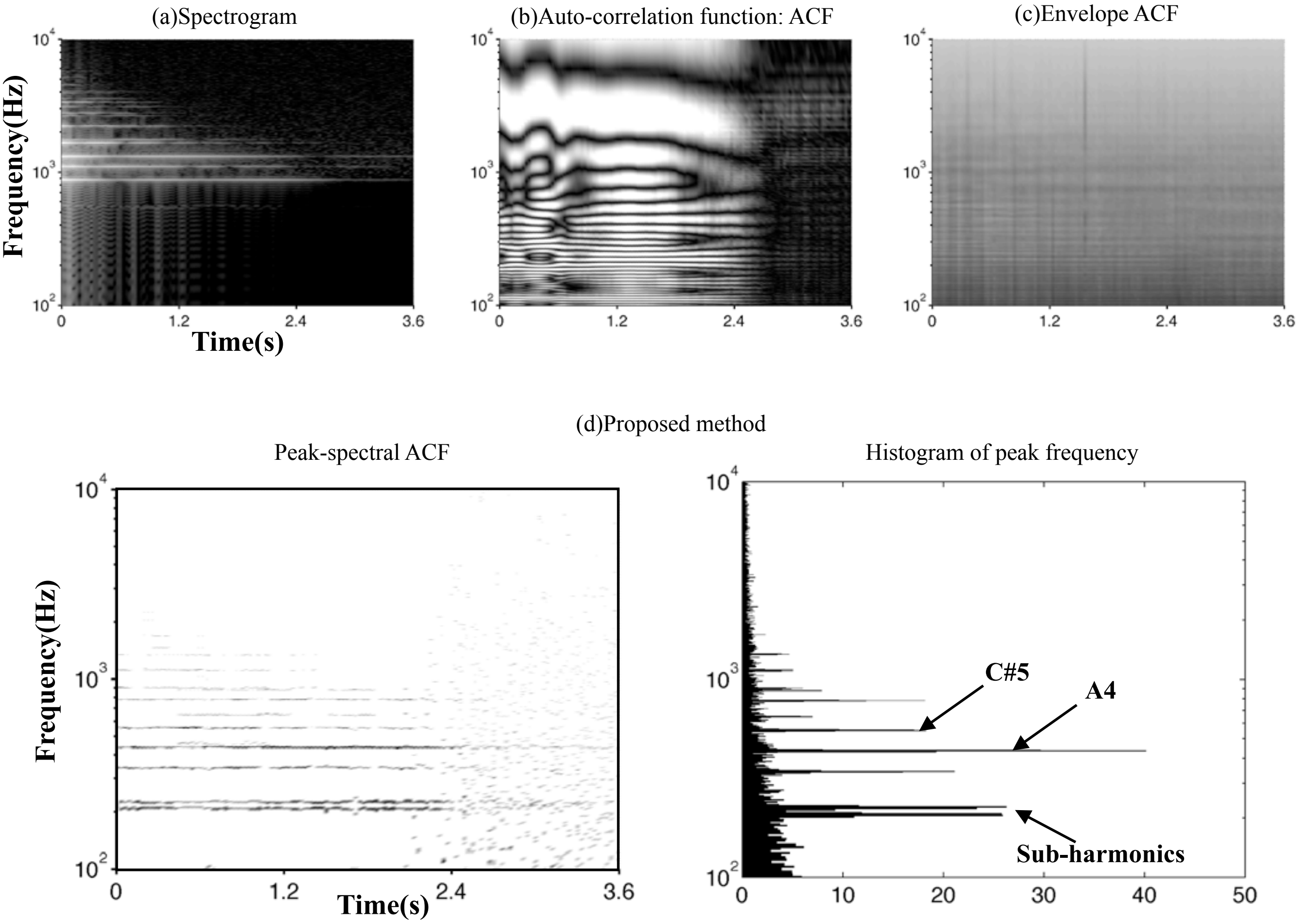
(c)Envelope ACF

(d)Proposed method

Peak-spectral ACF

Histogram of peak frequency

Fig. 6. Examples of fundamental frequency analysis: Piano tone, A4(≅440(Hz))



(a)Spectrogram

(b)Auto-correlation function: ACF

(c)Envelope ACF

(d)Proposed method

Peak-spectral ACF

Histogram of peak frequency

Fig. 7. Examples of fundamental frequency analysis: Piano tone, A4(≅440(Hz))+C#5(≅554(Hz))

4. CONCLUSIONS

This article has developed a new method for fundamental frequency analysis for sound. It is well known there is a difficulty in estimating the fundamental frequency independent of power spectral properties of a signal and its frequency band. The proposed method represents power spectral properties of a signal by a sequence of unit pulses that denote only the locations(frequencies) of dominant spectral peaks without magnitude information. Therefore, in principle, spectral-magnitude-free estimation of the fundamental frequency is possible even under the missing fundamental conditions. The results for tones of piano seem promising, if the results are compared with conventional methods. Analysis of compound sounds like musical chord, and more over tracing temporal change of the fundamental frequencies of other musical sounds are future studies.

The authors would thank Mr. Mikio Tohyama(SV research associates) for his fruitful suggestions and discussions.

REFERENCES

[1] Hess, W., (1983). Pitch Determination of Speech Signals. Springer-Verlag, Beriline

[2] Itakura, F., and Saito, S., (1971). Speech Information Compression based on the Maximum Likelihood Spectral Estimation [in Japanese], J. Acoust. Soc. Jpn. 19(9), pp.17-26

[3] Markel, J., D., (1972). The SIFT algorithm for fundamental frequency estimation, IEEE Trans. Audio, Electroacoust., AU-20(5), pp.367-377

[4] de Cheveigné, A., and Kawahara, H., (2002). YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. 111. pp. 1917-1930

[5] Goto, M., (1999). F0 Estimation of Melody and Bass Lines in Real-world Musical Audio Signals, Information Processing Society of Japan (IPSJ), 99(68), pp.91-98

[6] Kazama, M., Yoshida, K., and Tohyama, M., (2003). Signal Representation Including Waveform Envelope by Clustered Line-Spectrum Modeling, J. Audio Eng. Soc. 51, pp.123-137

[7] Ray, M., and Lowel, O, (1997). A unitary model of pitch perception, J. Acoust. Soc. Am. 102(3), 1811-1820