Title	時間周波数領域の瞬時振幅・瞬時周波数を利用した基本周波数推定法の検討
Author(s)	山口，敏浩
Citation	
Issue Date	2019-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/16153
Rights	
Description	Supervisor:鵜木祐史，情報科学研究科，修士（情報科学）

Study on method for estimating fundamental frequency using instantaneous amplitude and frequency in the time-frequency domain

1510755 Toshihiro Yamaguchi

This study aims to discuss how to propose a method for robustly and precisely estimating fundamental frequency (F0) of speech under noisy or reverberant conditions. The F0 of speech can be utilized as a significant feature to represent the source information (glottal waveform) of speech sound in various speech-signal processes. These are in speech analysis/synthesis systems, automatic speech recognition (ASR) systems, and speech emphasis methods. Therefore, a particularly important issue in these applications is to robustly and accurately estimate the F0 of target speech in real environments. F0 estimation is well-known to be one of the studies that is considered extremely difficult, for the following two reasons.

The first reason is that F0 cannot be observed directly. The F0 in humans is the frequency of vowels, which is a physical quantity defined as the reciprocal of the period of vocal cord vibration. In order to obtain F0, it is necessary to accurately estimate the vibration waveform of the vocal cords from the observed waveform, because human vocal cords cannot be observed directly. And pure estimation of vocal cord vibration can only be done by removing the influence of the vocal tract (formant frequency and spectral envelope) in some way, since the vocal cord vibration is recorded in the sound file via the vocal tract. Even if it is estimated, the accuracy is expected to decrease, because noise and reverberation are mixed with the observed waveform to be estimated.

The second reason is that the vocal cord vibration that is the source of F0 is quasi-periodic instead of a constant periodic vibration. In other words, if the temporal trajectory of F0 cannot estimate, the estimation is in vain. For these reasons, the F0 estimation requires robustness and accuracy.

Despite the fact that the F0 estimation of speech is one of the studies that have been conducted since ancient times, an estimation method that has both robustness and accuracy has not been established yet. The F0 estimation method with high performance has been proposed recently, which realized either robustness or accuracy. For example, those representative examples are FreeDAM (Fundamental fRequency Estimation mEthod using Demodulation of Amplitude Modulation) with high robustness, and a method with high accuracy proposed by Dhiman et al. FreeDAM is an F0 estimation method inspired by the concept of human's pitch perception called "Missing fundamental" and Modulation Transfer Function (MTF). The missing fundamental is a phenomenon in which the human perceived pitch does not change, and does not depend on the presence or absence of the complex sound. The concept of MTF is to show that the periodicity of F0 is maintained even under the influence of noise and reverberation. That implies that F0 can be estimated by applying AM demodulation technology, assuming that harmonic structure of sound is in the spectrum of AM sound.

The method proposed by Dhiman et al. Is a method to extract temporal variation of F0 smoothly from a sound spectrogram. The Riesz transform is used to transform a sound spectrogram into instantaneous information. Since this instantaneous information is a complex signal composed of an amplitude term and a phase term, instantaneous amplitude (IA) and instantaneous frequency (IF) is decomposed from each term. Dhiman, and others, estimate the smooth temporal variation of vocal cord sound source by using the property of this CRT. However, both FreeDAM with its focus on robustness, and the method proposed by Dhiman with high accuracy have their own problems. From this reason, the aim of this research is to make a basic study to solve each other's problems in that way, which each other's advantages strengthen each other's problems. That can realize the F0 estimation method that has both robustness and accuracy. First, we will deepen our knowledge on previous research and pursue measures to solve the problems. Then, the principle is considered for establishing a practical estimation method by the implementation and evaluation of the algorithm by computer. The background of this research is the formation of natural speech communication between human and machine. At present, machines have been equipped with artificial intelligence and communication functions, and it has become a reality in recent years that communication is being carried out mutually. It limited only simple matter, but, machine can easy communications with human at hotel or some shops in japan already. At the present time, the communication is limited to the simplicity of use, and it is not a nature like communication among people. In order to refine this machine and the person's communication, it is necessary to provide the machine with the function of combining naturalness with clarifying what is also naturalness. However, its naturalness is not yet fully understood.

Prosodic information such as intonation and accent included in speech is deeply related to the features that form the naturalness of human communication. This prosody combines the three attributes of pitch, loudness, and timber and from the synergistic effect, humans perceive communication as natural. However, one thing that cannot be forgotten is that the pitch is a psychological attribute and a sensory quantity. The evaluation is made subjectively, because each person ultimately perceives pitch differently. It is pointed out that pitch is causal between the pitch and the fundamental frequency of the sound. In other words, it suggests that the possibility of quantitative understanding of human senses is hidden in F0. For this reason, F0 estimation is an indispensable estimation technique for applications such as speech recognition, speech synthesis, voice quality evaluation, and sound source separation, and makes objective evaluation measures.

The results obtained in this study contribute to providing elemental technologies to realize natural communication between human and machine.