INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Robust method for estimating F0 of complex tone based on pitch perception of amplitude modulated signal

Kenichiro Miwa and *Masashi Unoki*

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
{threerings,unoki}@jaist.ac.jp

Abstract

Estimating the fundamental frequency (F0) of a target sound in noisy reverberant environments is a challenging issue in not only sound analysis/synthesis but also sound enhancement. This paper proposes a method for robustly and accurately estimating the F0 of a time-variant complex tone on the basis of an amplitude modulation/demodulation technique. It is based on the mechanism of the pitch perception of amplitude modulated signal and the frame-work of power envelope restoration based on the concept of modulation transfer function. Computer simulations were carried out to discuss feasibility of the accuracy and robustness of the proposed method for estimating the F0 in heavy noisy reverberant environments. The comparative results revealed that the percentage correct rates of the estimated F0s using five recent methods (TEMPO2, YIN, PHIA, CmpCep, and SWIPE') were drastically reduced as the SNR decreased and the reverberation time increased. The results also demonstrated that the proposed method robustly and accurately estimated the F_0 in both heavy noisy and reverberant environments.

Index Terms: F0 estimation, pitch perception, complex tone, amplitude modulation/demodulation, noisy reverberant environments

1. Introduction

Estimating the fundamental frequency (F_0) of a target sound is a significant challenge in not only sound analysis/synthesis but also various types of multimedia signal processing. For example, the F_0 is used as a factor to control the pitch of sound in sound analysis/synthesis and a significant cue for sound enhancement. Therefore, the F_0 needs to be robustly and accurately estimated from observed signals in real environments.

Estimation of the F_0 of a target sound has been widely studied in the literature on speech signal processing, and many methods have been proposed over the last half century [1]. The conventional methods can be divided into processing in the time domain or frequency domain, or both. Most of these methods use periodic features in the time domain (e.g., autocorrelation) or harmonic features in the frequency domain (e.g., comb filtering) [1]. The others suppress the vocal tract filter effect based on the source-filter model [1].

Some methods that precisely estimate the F_0 of target noiseless speech have been established (e.g., TEMPO [2] (TEMPO2 [3]) and YIN [4]) by comparing electro-glottal-graph information. It has been reported that both methods can be used to estimate the F_0 of target noiseless speech extremely accurately. However, they cannot precisely estimate F_0 in noisy environments [5], reverberant environments [6], or both [7].

Other methods have been proposed that robustly estimate the F_0 of observed speech in noisy environments [5, 8]. The instantaneous amplitude (IA) of speech has been reported to have fine harmonic features that are robust against background noise. The instantaneous frequency (IF) of speech has also been used to estimate F_0s accurately, but when used in TEMPO, their stability is sensitive to noise. More robust methods using the IF by using bandwidth equations with harmonicity [8] or using the periodicity and harmonicity [5] related to the IA and IF have been proposed.

On the other hand, these methods to estimate the F_0 of a target signal robustly and accurately have been studied in reverberant environments [6]. It has been revealed that the correctness of the F_0 estimated by using these methods was drastically reduced as the reverberation time increased while the F_0 estimated with periodicity and/or harmonicity on the complex cepstrum by using the source-filter model [6] was relatively robust and accurate. However, it has not yet been clarified whether all these methods can precisely estimate F_0 in very noisy reverberant environments.

The possibility of estimating the F_0 of a steady complex tone in noisy reverberant environments based on the pitch perception of AM tone has been studied [7, 9]. The results of computer simulations showed that our method could robustly and accurately estimate F_0 from a noisy reverberant tone. However the target signal was limited to only steady complex tone with a time-invariant F_0 in the long-term. This paper proposes a method for robustly and accurately estimating the F_0 of a time-variant target signal in very noisy reverberant environments.

2. Signal definitions, problems, and concept

2.1. Definitions and problems

A time-variant harmonic signal, $x(t)$, as a target signal, can be represented as a form of analytical signal:

$$x(t) = \sum_{k \in K} a_k(t) \exp(j\omega_k(t)t + j\theta_k(t)), \qquad (1)$$

where $a_k(t)$ is the instantaneous amplitude and $\theta_k(t)$ is the instantaneous phase. Here, k denotes the harmonic index and K is the number of harmonics ($k = 1, 2, \cdots, K$). Since $\omega_k(t) = 2\pi k F_0(t)$, the fundamental frequency, $F_0(t)$, is an instantaneous frequency so that it should be extracted from $x(t)$ using instantaneous cues. The task of estimating F_0 in noisy reverberant environments is to extract $F_0(t)$ from noisy reverberant speech signal $y(t)$.

$$y(t) = x(t) * h(t) + n(t), \qquad (2)$$

where $h(t)$ is the room impulse response (RIR) and $n(t)$ is the background noise. Operation "*" indicates the convolution.

http://dx.doi.org/10.21437/Interspeech.2017-1061

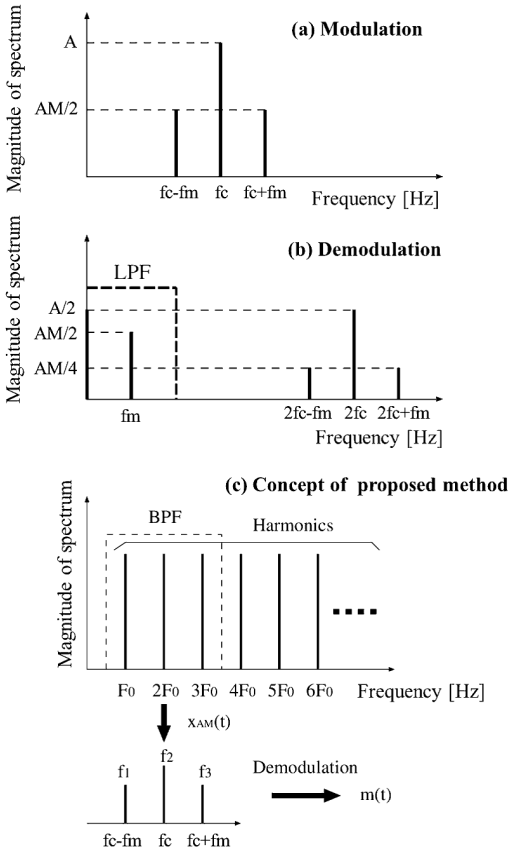

Figure 1: *Concept of F0 estimation based on pitch perception of AM signal: spectrum configuration of (a) amplitude modulation, (b) amplitude demodulation, and (c) F0 estimation.*

2.2. Amplitude modulation and demodulation

Amplitude modulation (AM) is a technique used in radio communication. The AM-signal fluctuates amplitude as the transmitted signal in relation to the message being sent on the carrier signal. The AM-signal, $x_{\mathrm{AM}}(t)$, is formulated as

$$x_{\mathrm{AM}}(t) = A\left\{1 + Mm(t)\right\}c(t). \tag{3}$$

Here, $m(t) = \cos(\omega_m t)$ and $c(t) = \cos(\omega_c t)$ are referred to as "message" and "carrier" signals, respectively, where $\omega_m = 2\pi f_m$ and $\omega_c = 2\pi f_c$. f_m and f_c are the modulation frequency and carrier frequency, respectively. A and M, are the carrier amplitude and magnitude of the message, respectively. Thus, M is referred to as the "modulation index".

By using trigonometric identities, it can be rewritten as

$$\begin{aligned} x_{\mathrm{AM}}(t) &= \frac{AM}{2}\left\{\cos((\omega_c - \omega_m)t) + \cos((\omega_c + \omega_m)t)\right\} \\ &\quad + A\cos(\omega_c t), \end{aligned} \tag{4}$$

where ω_c must be much greater than ω_m ($\omega_c \gg \omega_m$). Therefore, $x_{\mathrm{AM}}(t)$ has three components: a carrier (third term) and two sideband signals (first two terms), as shown in Fig. 1 (a).

The simplest form of AM is product detection. By multiplying the same carrier, we obtained

$$\begin{aligned} & x_{\mathrm{AM}}(t)\cos(\omega_c t) \\ &= \frac{AM}{4}\left\{\cos((2\omega_c - \omega_m)t) + \cos((2\omega_c + \omega_m)t)\right\} \\ &\quad + \frac{A}{2}\cos(2\omega_c t) + \frac{AM}{2}\cos(\omega_m t) + \frac{A}{2}. \end{aligned} \tag{5}$$

Then, as shown in Fig. 1 (b), by low-pass filtering with the cut-off frequency of $2\omega_m$ and by multiplying $2/AM$, we obtained

$$m(t) = \frac{2}{AM}\left(\mathrm{LPF}[x_{\mathrm{AM}}(t)\cos(\omega_c t)] - \frac{A}{2}\right) = \cos(\omega_m t), \tag{6}$$

where $\mathrm{LPF}[\cdot]$ is low-pass filtering.

2.3. Pitch perception of AM tone

People can easily perceive the pitch of a periodic complex tone (target signal) in realistic environments. In particular, we can perceive the pitch corresponding to the fundamental period $(1/F_0)$ of vibration, even if no energy is present at the fundamental. This phenomenon is referred to as "the missing fundamental" and has been widely studied. The pitch perception of AM tone is related to field of literatures. This phenomenon can be interpreted as the fundamental frequency relating to the differences between the center and side-lobe frequencies and/or the fundamental period observed in the temporal-lobe in the temporal amplitude are important cues for pitch perception [10].

2.4. Concept of proposed method

Accurately and robustly estimating the F_0 in computer systems is very difficult in very noisy reverberant environments while people can easily do the same task. Our concept for solving this problem is motivated from the above knowledge with regard to the pitch perception of the AM signal. Figure 1 (c) shows the harmonics of the complex tone with the F_0. Our concept is to extract any three harmonics from the complex tone and then demodulate the message from them in which the harmonic frequencies, f_1, f_2, and f_3, correspond to the sidebands and center frequencies in theAM domain, $f_c - f_m$, $f_c + f_m$, and f_c.

The extracted message has a periodic signal with $1/F_0$ so that the final step is to extract this fundamental period or the frequency of $m(t)$ as the estimated F_0. However, $m(t)$ is irregularly smeared due to the effects of noise and reverberation. Thus, envelope restoration based on the modulation transfer function (MTF) concept [12] is used to restore $m(t)$ as a maximization of the modulation index to be 1.0 by using the following equation.

$$E_x(z) = E_y(z)\left(1 - \exp\left(\frac{13.8}{\hat{T_R} f_s}\right) z^{-1}\right). \tag{7}$$

Here, $E_x(z)$ is the restored $m(t)$, $E_y(z)$ is the extracted $m(t)$ and f_s is the sampling frequency. $\hat{T_R}$ is controlled so that the modulation index of $m(t)$ is 1.0. The concept of our proposed method is to estimate F_0 robustly and accurately from the restored $m(t)$ in noisy reverberant environments.

3. Proposed method

Figure 2 explains the algorithm for estimating F_0 on the basis of our concept. This method is composed of five parts.

(0) **Assumption.** The target signal $x(t)$ is a time-variant complex tone and $F_0(t)$ is constant for each segment.

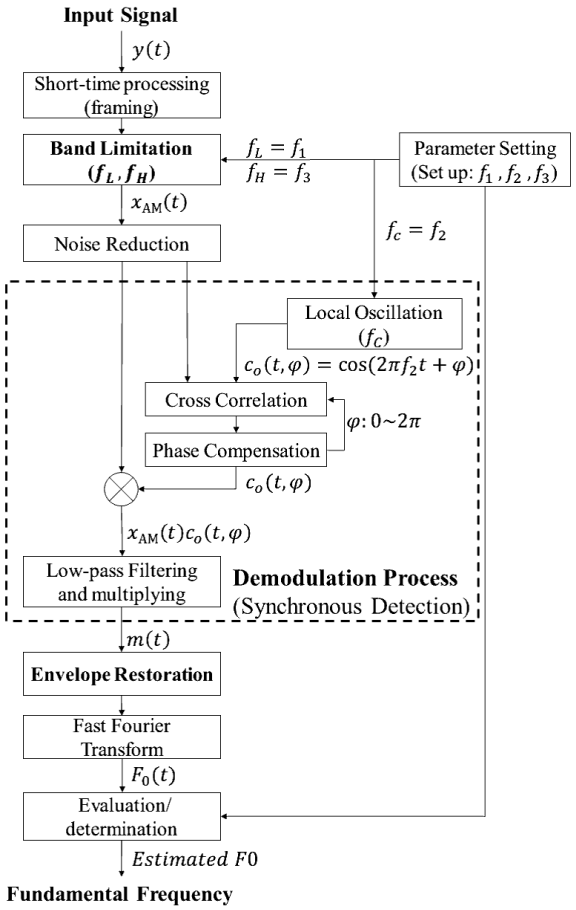

Figure 2: *Block diagram of proposed method.*

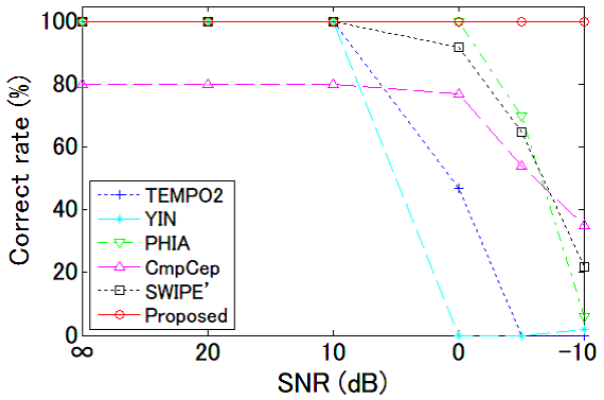

Figure 3: *Percent correct rate within error margin of* 5 *% for F0 estimation in noisy environments.*

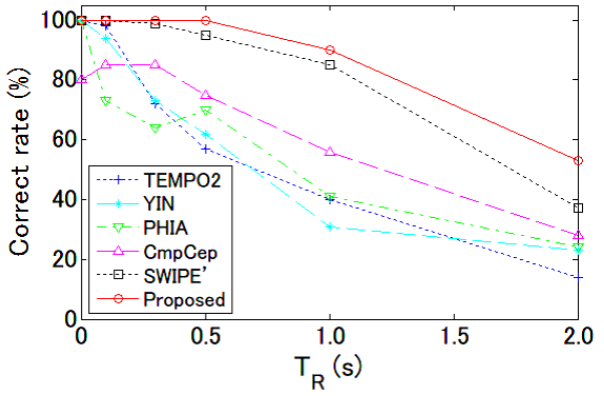

Figure 4: *Percent correct rate within error margin of* 5 *% for F0 estimation in reverberant environments.*

(1) **Band-limitation.** The band-pass filter (BPF) is used to extract all three harmonics (f_1, f_2, and f_3) as an AM signal, $x_{AM}(t)$, in Eq. (4) from $y(t)$. In this case, the lower frequency f_L and the upper frequency f_H in the band-limitation are set under $f_c - f_m$ ($= f_1$) and over $f_c + f_m$ ($= f_3$) in which the center frequency $f_c = f_2$. In the algorithm, a candidate of f_2 is set as $f_2 = 2F_0$ and then f_1 and f_3 are determined as $f_1 = F_0$ and $f_3 = 3F_0$.

(2) **Noise reduction.**

General spectral subtraction is used to remove inharmonic noise under the averaged spectra from the output of the BPF.

(3) **Demodulation process.** Product detection is used as the amplitude demodulator to extract $m(t)$ from the three extracted harmonics above. In this case, the auto-correlation method is used to synchronize the cosine wave as a local oscillator into $\cos(\omega_c t)$ in Eq. (5). This process is phase compensation. Then, by using LPF, $m(t)$ can be extracted from $x_{\mathrm{AM}}(t)$.

(4) **Envelope Restoration.** By using Eq. (7), $m(t)$ is restored so that the modulation index M can reach 1.0.

(5) F_0 **determination.** Period $1/f_m$ in $m(t)$ can be derived by using the auto-correlation function of $m(t)$ and/or spectrum analysis by fast Fourier transform (FFT). If f_m is completely identified as the candidate for F_0, the estimated f_m is the estimated fundamental frequency $\hat{F}_0$. These blocks of the demodulation process and F_0 determination run with the number of search candidates of F_0. The identified F_0 is the estimated $F_0(t)$ in the proposed method.

4. Evaluations

We considered whether the proposed method is applicable to a time-variant complex tone in very noisy reverberant environments. Thus, computer simulations were carried out to evaluate how the proposed method robustly and accurately estimates the F_0 in very noisy or/and reverberant environments. We compared with the five typical methods: TEMPO2 [3], YIN [4], PHIA [5], Complex Cepstrum (CmpCep) [6] and SWIPE' [14].

The target signal $x(t)$ in these simulations was a time-variant complex tone with 10 harmonics ($K = 10$) in which the F_0 in the tone at every 250-ms segment varied from 100 to 200 Hz with an F_0 trajectory of 100, 125, 150, 175, 200, 180, 160, 140, 120 and 100 Hz. Background noises, $n(t)$s, were white Gaussian noises with SNRs of 20, 10, 0, -5 and -10 dB. 10 different $n(t)$s were generated at each SNR condition. RIRs, $h(t)$s, were Schroeder's RIRs and were formulated as

$$h(t) = a \exp\left(-6.9t/T_R\right) c(t), \tag{8}$$

where a is the gain factor as the normalized power of $h(t)$, T_R is the reverberation time, and $c(t)$ is the white noise carrier. This is the well-known stochastic approximation of RIR [13]. Five reverberation conditions ($T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s) were used. 10 different RIRs, $h(t)$s, were generated by 10 different white noise carriers $c(t)$s at each T_R condition.

All the stimuli that we used in these simulations were noisy or/and reverberant signals. There were 50 stimuli under noisy

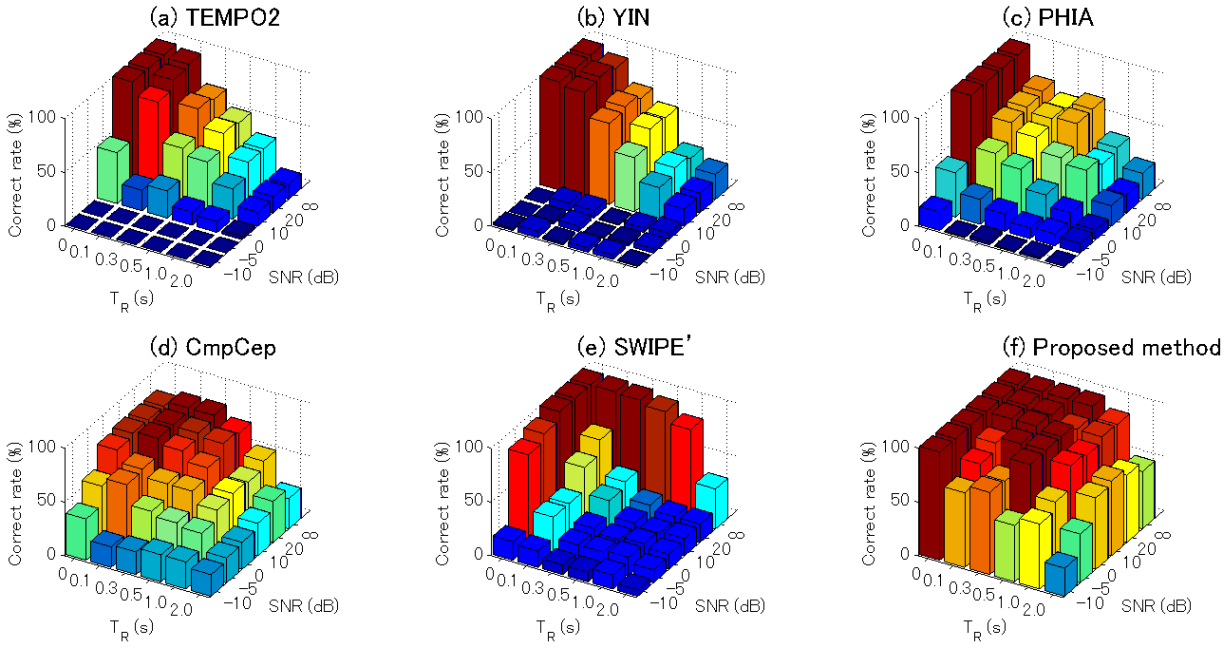

Figure 5: *Percent correct rate within error margin of* 5 *% for F0 estimation in noisy reverberant environments.*

conditions ($y(t) = x(t) + n(t)$), 50 stimuli under reverberant conditions ($y(t) = x(t) * h(t)$, and 250 stimuli under noisy reverberant conditions ($y(t) = x(t) * h(t) + n(t)$).

We used the percent correct rate (%), which is defined as

$$\text{Correct rate} = \frac{N_{F_{0,\text{Est}}}(E)}{N_{F_{0,\text{Ref}}}} \times 100, \tag{9}$$

where $F_{0,\text{Ref}}(t)$ and $F_{0,\text{Est}}(t)$ are the reference F_0 and estimated F_0. $N_{F_{0,\text{Est}}}(E)$ is the size of the correct region that satisfies $|F_{0,\text{Ref}}(t) - F_{0,\text{Est}}(t))|/F_{0,\text{Ref}}(t) \leq E(\%)$ within the voiced section (t) where E is the error margin (%). $N_{F_{0,\text{Ref}}}$ is the size of region $F_{0,\text{Ref}}(t)$ in the voiced section. In this paper, the F_0 set up in the original signal is used as the correct F_0 (reference F_0, $F_{0,\text{Ref}}(t)$). $F_{0,\text{Est}}(t)$ was used to estimate F_0 with the six methods in noisy reverberant environments. Here, $E = 5$ % was used in the evaluation.

Figure 3 shows the results of the comparative evaluations for the proposed methods and the five typical methods for estimating F_0 in noisy environments, as a function of SNR. Most of the methods are robust against noise under low-noise conditions. Although the correct rates of the five methods are reduced as the SNR decreases, the correct rate of the proposed method is still high (100 %).

Figure 4 shows the results of the comparative evaluations for all methods for estimating F_0 in reverberant environments, as a function of T_R. Although the correct rate of TEMPO2, YIN, PHIA and CmpCep are smoothly reduced as the T_R increased, the correct rate of SWIPE' and the proposed method are still high under short reverberant conditions. Under all conditions, the proposed method is the most robust against reverberation.

Figure 5 shows the total results of the comparative evaluations for all methods that we used for estimating F_0 in noisy reverberant environments. TEMPO2 and YIN provide high performance under noiseless condition. Although PHIA is robust for noisy environments, the correct rate is reduced as T_R increases. CmpCep is robust against noise and reverberation; however, it is not an accurate estimation. SWIPE' provides high performance in either noisy or reverberant environments, but the correct rate is drastically reduced in both environments. From all results, it is shown that the proposed method can robustly and accurately estimate F_0 in very noisy reverberant environments.

5. Conclusion

This paper proposed a method for robustly and accurately estimating the F_0 of a time-variant complex tone in very noisy reverberant environments on the basis of the pitch perception of AM tone. Besides, the envelope restoration function based on the MTF concept was applied to the proposed method. Computer simulations were carried out to evaluate the robustness and accuracy of the proposed method compared to those of the other typical methods (TEMPO2, YIN, PHIA, CmpCep and SWIPE'). The results revealed that the typical methods could not robustly or accurately estimate the F_0 in very noisy reverberant environments. The results also demonstrated that the proposed method is robust against heavy noise as well as long reverberation and can estimate the F_0 from the observed signal. For future work, we plan to further develop our method to ensure even more robust and accurate estimates of the F_0 for real signals such as the sounds of instruments or speech.

6. Acknowledgements

This work was supported by a Grant-in-Aid for challenging Exploratory Research (No. 16K12458) and Innovative Areas (No. 16H01669) from MEXT, Japan, and by the Secom Science and Technology Foundation.

7. References

[1] W. J. Hess, "Pitch and Voicing Determination," in Advances in speech signal processing, Eds. S. Furui and M. M. Sondhi, 3–48, Marcel Dekker. Inc. New York, 1992.

[2] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a reptitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[3] Kawahara, H., Katayose, H., Cheveigné, A., and Patterson, R. D., "Fixed Point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," *Proc. Eurospeech'99*, vol. 6, pp. 2781–2784, 1999.

[4] Cheveigné, A., Kawahara, H., "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[5] Ishimoto, Y., Unoki, M., and Akagi, M., "A Fundamental Frequency Estimation Method for Noisy Speech Based on Instantaneous Amplitude and Frequency," *Proc. EuroSpeech2001*, pp. 2439–2442, 2001.

[6] Unoki, M., Hosorogiya, T., and Ishimoto, Y., "Comparative evaluations of robust and accurate F0 estimates in reverberant environments," *Proc. ICASSP2008*, pp. 4569–4572, 2007.

[7] K. Miwa and M. Unoki, "Study on Method for estimating F0 of steady complex tone in noisy reverberant environments," *Proc. IIH-MSP2013*, pp. 456–459, Oct. 2013.

[8] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," *Proc. ICSLP2000*, no. 2, pp. 907–910, 2000.

[9] K. Miwa and M. Unoki, "A Method of Estimating Fundamental Frequency of Comlex Tone Based on Pitch Perception of AM Signal," *IEICE Trans.* Vol. J98-A, No. 12, pp. 668–679, 2015 (written in Japanese with English abstract).

[10] R. J. Zatorre, "Pitch perception of complex tones and human temporal-lobe function," *J. Acoust. Soc. Am.*, vol. 84, no. 2, pp. 566–572, 1988.

[11] Unoki, M., Sakata, K., Furukawa, M., and Akagi, M., "A speech dereverberation method based on the MTF concept in power envelope restoration," *Acoust. Sci. &Tech.*, vol. 25, no. 4, pp. 243–254, 2004.

[12] Unoki, M., Yamasaki, Y., Akagi, M., "MTF-based power envelope restration in noisy reverberant environments," *Proc. EUSIPICO2009*, pp. 228–232, 2009.

[13] Schroeder, M. R., "Modulation Transfer Functions: Definition and Measurement," *Acustica*, Vol. 49, pp. 179–182, 1981.

[14] Camacho, A., Harris, J. G. "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol.124, no. 3, pp. 1638–1652, 2008.