COMPARATIVE EVALUATIONS OF ROBUST AND ACCURATE F0 ESTIMATES IN REVERBERANT ENVIRONMENTS

Masashi Unoki[(1)]*, Toshihiro Hosorogiya*[(1)]*, and Yuichi Ishimoto*[(2)]

[(1)]School of Information Science, Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi, Ishikawa 923-1292, JAPAN
[(2)] School of Media Science, Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo 192-0982, JAPAN

ABSTRACT

This paper reports comparative evaluations of the method we previously proposed of estimating fundamental frequency (F_0) based on complex cepstrum analysis with nine typical methods over huge speech-sound datasets in both artificial and realistic reverberant environments (in room acoustics). They involve several classic algorithms (Cepstrum, AMDF, LPC, and modified autocorrelation) and a few modern algorithms (TEMPO, YIN, and PHIA). The comparative results revealed that the percentage correct rates of the estimated F0s using them were drastically reduced as the reverberation time increased while F_0 estimated with the proposed method was completely robust and accurate. They also demonstrated that homomorphic analysis and the concept of a source-filter model were relatively effective for estimating F_0. The results also demonstrated that it was much better than the previously reported methods in terms of robustness and providing accurate F_0 estimates in both artificial and realistic reverberant environments.

Index Terms— F0 estimation, reverberant speech, complex cepstrum analysis, MTF concept, source-filter model

1. INTRODUCTION

The fundamental frequency (F_0) 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 F_0 of target speech in real environments.

Many studies on estimating the F_0 of target speech have been done in the literature on speech-signal processing, and many methods have been proposed [1] over the last half century. The traditional methods of estimation can be divided into processing in the time and frequency domains, or both. Most of these have made use of the periodic features of speech in the time domain (e.g., autocorrelation) or harmonic features in the frequency domain (e.g., comb filtering) [1]. However, the problem of estimating the F_0 still seems to be unsatisfactorily resolved because three main issues remain, i.e., (i) **observability**: the observed speech is an emitted sound passing through the mouth/nose so that it is impossible to directly observe glottal vibrations from it, (ii) **flexibility and irregularity**: glottal vibrations are not complete periodic signals and the range of variations in the periods is relatively wide, and (iii) **robustness**: the observed speech signals are affected by noise and reverberation so that significant features for estimating F_0 are also smeared.

Most studies have focused on the first two issues so that they have implicitly assumed all speech signals are observed in clean environments or all observations are only noiseless speech sounds. Various methods of estimating F_0 have been proposed under this assumption to solve the first issue by suppressing the effects of filter characteristics, based on the source-filter model, from the observed speech sounds (e.g., homomorphic-analysis [2] and LPC methods [1]). A few approaches to precisely estimate the F_0 of target noiseless speech have been established (e.g., TEMPO [3] and YIN [4]) by comparing electro-glottal-graph (EGG) information. It has been reported that both methods can be used to estimate the F_0 of target noiseless speech extremely accurately so that the first two issues seem to be resolved. However, it has not yet been clarified whether these methods can precisely estimate F_0 in real environments.

It is generally known that the method of estimating F_0 using periodic and/or harmonic features is relatively robust against background noise [1, 5, 6]. Moreover, it has been reported that the instantaneous amplitude (IA) of speech has fine harmonic features that are robust against background noise. The instantaneous frequency (IF) of speech has also been used to accurately estimate F_0s but their stability as used in TEMPO is sensitive to noise. More robust methods using IF have been proposed by using bandwidth equations with harmonicity [5] or using periodicity and harmonicity [6] related IA and IF. It has been reported that these are more robust than TEMPO and can precisely estimate the F_0 in noisy environments.

All these methods have focused on noiseless to noisy conditions to estimate sufficiently accurate F_0s of target speech. Thus, methods using IA and IF or those with robust features against noise such as periodicity and harmonicity have been regarded as accurately being able to estimate F_0s from noisy speech. The last issue seems to be have been solved at this time; however, there have been no studies on robustness in realistic reverberant environments. In our study on simulations [7], we found that no typical methods worked as well in artificial reverberant environments and their percentage correct rates for F_0s were reduced drastically as reverberation time increased. We thus proposed a method of estimating F_0 from reverberant speech by utilizing the MTF concept and the source-filter model in complex cepstrum analysis [7]. However, the method then proposed was only evaluated in artificial reverberant environments (stochastic approximation). In this paper, our aim is to compare our evaluations of our latest method of estimating F_0 with traditional methods in terms of robustness and accuracy in realistic reverberant environments (e.g., concert hall, lecture room, and church) to clarify the last issue.

This work was supported by a Grant-in-Aid for Scientific Research (No. 18680017) from the Ministry of Education, Japan. It was also partially supported by the SCOPE (071705001) of MIC, Japan.

2. PROPOSED METHOD

2.1. Problem with estimating F_0

A time-varying harmonic signal, $x(t)$, can be represented as

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

where $a_k(t)$ is the instantaneous amplitude and $\theta_k(t)$ is the phase. Here, k denotes the harmonic index and K is the number of harmonics. Since $\omega_k(t) = 2\pi k F_0(t)$, the fundamental frequency, $F_0(t)$, is an instantaneous frequency so that this should be extracted from $x(t)$ using instantaneous cues. The task of estimating F_0 in reverberant environments is to extract $F_0(t)$ from reverberant speech signal $y(t)$ or the respective short-term Fourier transform (STFT), $Y(\omega, \tau)$:

$$y(t) = x(t) * h(t) = e(t) * v_\tau(t) * h(t), \quad (2)$$
$$Y(\omega, \tau) = X(\omega, \tau)H(\omega, \tau) = S(\omega, \tau)V(\omega, \tau)H(\omega, \tau), \quad (3)$$

where $X(\omega, \tau)$ and $H(\omega, \tau)$ are the STFTs of $x(t)$ and $h(t)$ in room acoustics. The $e(t)$ is the source signal related to glottal information and $v_\tau(t)$ is the impulse response of the filter related to the vocal tract at time τ. $S(\omega, \tau)$ is the STFT of $e(t)$ and $V(\omega, \tau)$ is that of $v_\tau(t)$. Note that $H(\omega, \tau)$ is actually required to present all characteristics of $h(t)$ by using a long-term Fourier transform (LTFT) so that analysis length should take longer than the reverberation time.

2.2. Complex cepstrum analysis

From Eq. (3), the complex cepstrum of $y(t)$ can be represented as

$$C_Y(q, \tau) = C_{\mathrm{src}}(q, \tau) + C_{\mathrm{flt}}(q, \tau) + C_H(q, \tau), \quad (4)$$

where $C_H(q, \tau)$ is the complex cepstrum of the reverberant impulse response, $h(t)$. $C_{\mathrm{src}}(q, \tau)$ and $C_{\mathrm{flt}}(q, \tau)$ are the complex cepstra of source and filter characteristics. These cepstra can also be represented as all amplitude and phase cepstra (denoted by subscripts "A" and "ϕ"). The complex cepstrum can also be separately represented as minimum and non-minimum phase characteristics (denoted by subscripts "min" and "all"). Since $|X_{\mathrm{all}}(\omega, \tau)| = 1$ and $C_{A,\mathrm{all}}(q, \tau) = 0$, $C_Y(q, \tau)$ can be separately represented as

$$\begin{aligned} &C_{Y,A,\mathrm{min}}(q, \tau) + C_{Y,\phi,\mathrm{min}}(q, \tau) + C_{Y,\phi,\mathrm{all}}(q, \tau) \\ &= C_{\mathrm{src,A,min}}(q, \tau) + C_{\mathrm{src},\phi,\mathrm{min}}(q, \tau) + C_{\mathrm{src},\phi,\mathrm{all}}(q, \tau) \\ &\quad + C_{\mathrm{flt,A,min}}(q, \tau) + C_{\mathrm{flt},\phi,\mathrm{min}}(q, \tau) + C_{\mathrm{flt},\phi,\mathrm{all}}(q, \tau) \\ &\quad + C_{H,A,\mathrm{min}}(q, \tau) + C_{H,\phi,\mathrm{min}}(q, \tau) + C_{H,\phi,\mathrm{all}}(q, \tau). \quad (5) \end{aligned}$$

According to Eq. (4), an optimal F_0 estimate is only used to extract $C_{\mathrm{src}}(q, \tau)$ from $C_Y(q, \tau)$ to deal with the periodicity/harmonicity of the source information as a filter and the reverberation characteristics are eliminated. It is too difficult only to deal with $C_{\mathrm{src}}(q, \tau)$ in this task of estimation, without measuring $h(t)$ or $C_H(q, \tau)$. In addition, long-term $C_H(q, \tau)$, in which analysis takes longer than the reverberation time, is needed to accurately extract $C_{\mathrm{src}}(q, \tau)$.

2.3. Proposed method of estimating F_0

We found the following useful facts in our previous work [7]:

(1) The all-pass phase component of $h(t)$ can be regarded as a dominant effect from comparisons of robust and accurate F_0estimates. Therefore, $C_{H,\phi,\mathrm{all}}(q, \tau)$ can be canceled out in Eq. (5) by LTFT.

(2) Based on the modulation transfer function (MTF) concept, we can establish how much reverberation affects a reduction in the modulation index and we can then predict the characteristics of room acoustics (T_R) using inverse MTF. Therefore, we can easily predict that $C_{H,A}(q, \tau)$ will become a cepstral shape which exponential decay with respect to quefrency. The $h(t)$ can also be estimated by utilizing $\hat{a} \exp(-6.9t/\hat{T_R})$ with simulated white noise $\hat{n}(t)$ as a stochastic approximation (for estimating T_R in detail, see [7]).

(3) There is a Hilbert transform relationship between $C_{A,\mathrm{min}}(q, \tau)$ and $C_{\phi,\mathrm{min}}(q, \tau)$, and $C_{H,\phi,\mathrm{min}}(q, \tau)$ has the same characteristics in the positive quefrency domain based on the minimum phase characteristics. $C_{H,\mathrm{min}}(q, \tau)$ in the lower quefrency parts is generally larger than those in the higher parts and this attenuates exponentially as the quefrency increases. Therefore, $C_{H,\mathrm{min}}(q, \tau)$ have been assumed to concentrate in the lower quefrency parts.

(4) The cepstrum components of the source characteristics are separately concentrated in the higher quefrency parts and those of the filter are separately concentrated in the lower parts based on the advantages of the source-filter model. Therefore, if a component in the lower quefrency parts can only be removed by liftering, $C_{\mathrm{flt}}(q, \tau)$ and $C_{H,\mathrm{min}}(q, \tau)$ can be canceled out in Eq. (5).

The algorithm for estimating F_0 based on complex cepstrum analysis, the MTF concept, and the source-filter model are explained in Fig. 1. This method is composed of three main processes: (A) estimating the MTF-based reverberation impulse responses and eliminating them from reverberant speech, (B) extracting $X_{\mathrm{src}}(\omega, \tau)$ from the processed reverberant speech by using liftering based on the source-filter model, and (C) estimating F_0 from them by using a final decision block. Comb filtering was employed in the final block in Fig. 1. Lifter ($l(q) = 1, q > q_{\mathrm{lif}}$ and $l(q) = 0, q \leq q_{\mathrm{lif}}$) is used to cancel them out in Eq. (5). Here, $q_{\mathrm{lif}} = 1.25$ ms. This means the upper limit for estimating F_0 is 800 Hz. For details, see [7].

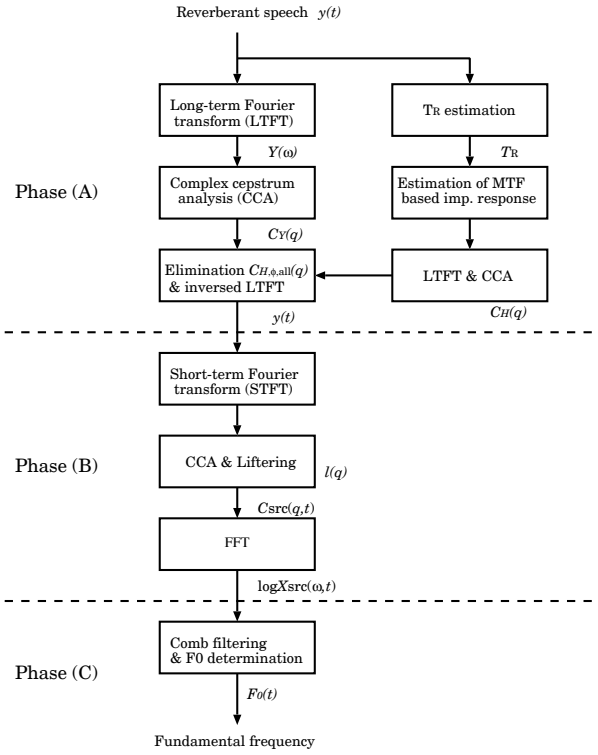

Fig. 1. Algorithm for proposed method.

Table 1. Comparison of percent correct rate (%) within error margin of 5 % in actual reverberant environments. IRdata corresponds to File No. in [8]. Reverberation time, T_R, is the average of T_Rs for transfer functions at 125 Hz to 8 kHz at octave frequencies. Bold and italic faces indicate best and worst results. RB, AB, and AC are "reflex board", "absorption board", and "absorption curtain".

Room condition (Impulse response)	IRdata	T_R (s)	TEMPO	YIN	Cepstrum	SrcFlt	Prop(Org)	Prop(Est)
Multi-purpose hall 1 with RB	301	1.09	*28.89*	35.51	42.22	40.62	**50.70**	44.99
Multi-purpose hall 1 without RB	302	0.80	*33.13*	38.36	47.54	47.04	**56.79**	51.67
Multi-purpose hall 4 with AB	307	1.42	*30.21*	38.21	46.89	49.52	**60.54**	54.50
Multi-purpose hall 4 without AB	308	1.54	*29.44*	37.21	46.04	49.01	**60.04**	54.20
Classic concert hall 1 ($d = 6$ m)	310	2.34	*29.61*	33.71	43.13	46.93	49.59	**50.37**
Classic concert hall 1 ($d = 11$ m)	311	2.35	*24.90*	29.85	37.51	41.02	**48.92**	45.66
Classic concert hall 1 ($d = 15$ m)	312	2.39	*18.26*	26.07	32.70	31.55	**40.72**	35.46
Classic concert hall 1 ($d = 19$ m)	313	2.38	*14.73*	22.97	29.68	27.32	**37.38**	29.76
Classic concert hall 2	314	1.14	*24.19*	32.23	38.12	36.11	**45.18**	40.51
Classic concert hall 4 with AC	316	1.92	*23.53*	31.26	38.46	38.64	**50.82**	42.81
Classic concert hall 4 without AC	317	2.55	*20.47*	27.23	34.49	36.13	**48.09**	40.33
Classic concert hall 5	323	2.32	*25.41*	33.39	41.95	41.79	**48.20**	45.70
Classic concert hall 6 (1F front)	324	1.77	*29.99*	36.07	45.29	47.20	**53.82**	51.61
Classic concert hall 6 (2F side)	325	1.74	*34.16*	38.28	47.13	49.79	**55.87**	54.07
Classic concert hall 6 (3F)	326	1.69	*18.38*	23.61	27.84	28.55	**41.62**	31.53
Lecture room with flatter echoes	201	1.36	*32.56*	41.51	53.50	51.48	**57.89**	55.36
Theater hall	318	0.85	*32.90*	38.06	46.39	45.05	**54.16**	50.09
Meeting room	401	0.62	*57.04*	55.28	70.26	70.25	**72.58**	71.14
Lecture room (capacity: 400 m^3)	402	1.12	*36.74*	47.03	61.52	56.74	**61.78**	60.14
Lecture room (capacity: 2, 400 m^3)	403	1.09	*26.48*	35.57	44.59	42.22	**52.71**	46.30
General speech hall (capacity: 11, 000 m^3)	404	1.54	*23.34*	31.97	40.04	38.11	**47.47**	41.71
Church 1 (capacity: 1, 200 m^3)	405	0.71	*32.46*	38.97	47.31	43.66	**52.27**	48.22
Church 2 (capacity: 3, 200 m^3)	406	1.30	*23.67*	30.32	36.84	36.15	**45.29**	41.91
Event hall 1 (capacity: 28, 000 m^3)	407	3.03	*16.99*	22.81	26.91	27.23	**37.68**	31.94
Event hall 2 (capacity: 41, 000 m^3)	408	3.62	*15.19*	21.78	26.38	27.14	**37.61**	29.68
Gym 1 (capacity: 12, 000 m^3)	409	2.82	*19.19*	25.95	31.25	32.81	**44.95**	35.07
Gym 2 (capacity: 29, 000 m^3)	410	1.70	*22.35*	27.77	31.70	32.67	**45.67**	36.08
Living room in wooden house	411	0.36	74.24	*65.35*	**81.45**	73.40	72.08	69.72
Movie theater	412	0.38	*42.88*	43.30	52.16	51.96	**59.32**	56.85
Concourse at train station	415	1.95	*20.89*	24.79	27.57	29.88	**44.71**	36.44

3. COMPARATIVE EVALUATIONS

3.1. Typical methods of estimating F_0

We evaluated nine typical methods to evaluate how robust estimates of F_0 were in reverberant environment. These were AMDF [1], STFT-ACorrLog (AutoCorrelation of Log-amplitude spectrum) [1], STFT-Comb (Comb filtering) [1], SHS (sub-harmonic summation) [1], Cepstrum [2], LPC-residue [1], TEMPO [3], YIN [4], and PHIA (Periodicity/Harmonicity using IA) [6]. Although other methods have been proposed, we chose these nine because they are commonly used in comparative evaluations and these others are just heavy revisions of them. We also evaluated the proposed method with (labeled "Prop(Est)") and without (labeled "Prop(Org)") T_R estimates. With and without comparisons of the proposed method were done to find how accurate the T_R estimates were. We compared them with typical methods, and a modified complex cepstrum method based on the source-filter model (labeled "SrcFlt"). The SrcFlt method was used to find how effectively $C_{H,\phi,\mathrm{all}}(q,\tau)$ was eliminated by the LTFT with the proposed method.

3.2. Sound dataset and evaluation measures

The sound dataset we used in this evaluation was the speech database of simultaneous recordings of speech and EGG by Atake *et al.* [5]. This dataset consisted of 30 short Japanese sentences uttered by 14 males and 14 females with voiced-unvoiced labels (total of 840 utterances, sampling frequency of 16 kHz, and quantization of 16 bits).

The reverberant speech sentences were created by convolving the original signals with the reverberant impulse responses, $h(t)$s.

$$h(t) = a \exp\left(-6.9t/T_R\right) n(t), \tag{6}$$

where a is a gain factor as the normalized power of $h(t)$, T_R is the reverberation time, and $n(t)$ is white noise. This is the well-known stochastic approximation impulse response in room acoustics [7]. This formulation for the impulse response has been used in a study on speech intelligibility in room acoustics as general artificial reverberation and thus has non-minimum phase components [7]. Six reverberation conditions ($T_R = 0.0, 0.1, 0.3, 0.5, 1.0$, and 2.0 s) were used in this study. There were a total of $5,040$ stimuli. Realistic-reverberant speech sentences were also created by convolving the original signals with 30-realistic reverberant impulse responses in room acoustics [8]. There were a total of $25,200$ stimuli.

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

$$\text{Correct rate} = \frac{N_{F_{0,\mathrm{Est}}}(E)}{N_{F_{0,\mathrm{Ref}}}} \times 100, \tag{7}$$

where $F_{0,\mathrm{Ref}}(t)$ and $F_{0,\mathrm{Est}}(t)$ are reference F_0 and estimated F_0. $N_{F_{0,\mathrm{Est}}}(E)$ is the size of the correct region that satisfies $|F_{0,\mathrm{Ref}}(t) -$

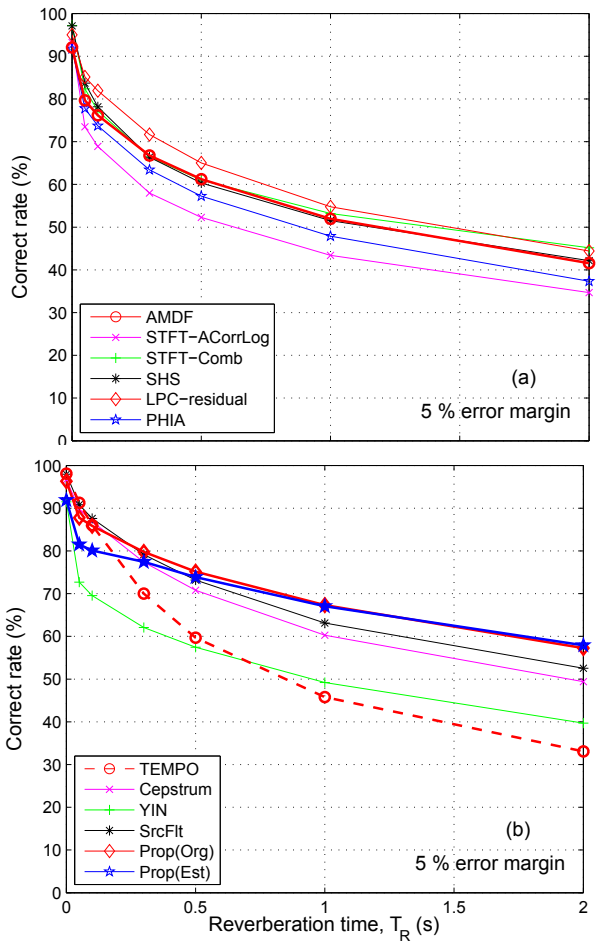

Fig. 2. Estimation results: percent correct rate within error margin of 5 % of F_0 estimates from reverberant speech as function of T_R.

$F_{0,\mathrm{Est}}(t))|/F_{0,\mathrm{Ref}}(t) \leq E(\%)$ within the voiced section (t) where E is the error margin (%). $N_{F_{0,\mathrm{Ref}}}$ is the size of region $F_{0,\mathrm{Ref}}(t)$ in the voiced section. In this paper, the F_0 estimated by TEMPO from the EGG signal is used as the correct F_0 (reference F_0, $F_{0,\mathrm{Ref}}(t)$). $F_{0,\mathrm{Est}}(t)$ was used to estimate F_0 with the twelve methods from reverberant speech signals. Here, $E = 5$ % was used in the evaluation.

3.3. Results

Figure 2 plots the results of comparative evaluations for the typical methods of estimating F_0 from reverberant speech as a function of T_R. This figure plots the percent correct rates for F_0 estimates. The correct rates of typical methods are drastically reduced as the reverberation time increases. The correct rates for typical methods were less than 50 % when T_R was 2.0 s. Although the overall accuracy of F_0 estimates tended to be reduced as reverberation time increased, about a 10 % improvement in the correct rates could be obtained with the new method. There is less difference in the results for both the proposed methods with and without T_R estimates. This means the T_R estimates can work as well. Since a correct rate of 60 % at $T_R = 2.0$ s, was achieved with the method we propose, we concluded that MTF-based impulse responses can be precisely estimated by utilizing T_R estimates. The results from the SrcFlt method indicate a small improvement (about 3 % in the correct rate) to that with the cepstrum method. In contrast, there were improvements of about 7 % in the percent correct rate by using the new method. We concluded that the use of complex cepstrum analysis with regard to non-minimum phase characteristics was effective for estimating F_0 in reverberant environments.

The proposed method and some other typical methods (TEMPO, YIN, Cepstrum, and SrcFlt) were compared and evaluated for realistic reverberant speech signals. Table 1 lists the results of estimates (averaged percent correct rate under all conditions). The results for the other methods are not listed here because there were no drastic improvements. Most of the results achieved by Prop(Org) were the best overall. This table indicates that Prop(Est) works almost as well as Prop(Org) when T_R is accurately estimated. Improvements achieved with the proposed methods were over 20 % with TEMPO and over 10 % with Cepstrum. This suggests that the proposed algorithms are alternatives for solving the last issue of robustness in terms of reverberation.

4. CONCLUSION

We evaluated the robustness and accuracy of twelve methods of estimating F_0 (i.e., classic AMDF, STFT-based, cepstrum, LPC, and SHS algorithms, modern PHIA, YIN, and TEMPO algorithms, and our proposed algorithms) in both artificial and realistic reverberant environments using huge speech datasets. The results revealed that none of the typical previously reported methods could accurately estimate F_0 in reverberant environments and that their accuracies drastically decreased as reverberation time increased. The results also demonstrated that periodicity and/or harmonicity on the complex cepstrum with the source-filter model concept and the MTF concept could effectively be used to estimate F_0 in reverberant environments. We demonstrated that our new method is robust against reverberation and can accurately estimate F_0 from observed reverberant speech.

5. REFERENCES

[1] 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] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, **41**(2), 293–309, 1966.

[3] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed Point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," *Proc. Eurospeech99*, **6**, 2781–2784, 1999.

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

[5] 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*, **2**, 907–910, 2000.

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

[7] M. Unoki, and T. Hosorogiya, "Estimation of fundamental frequency of reverberant speech by utilizing complex cepstrum analysis," J. Signal Processing, **12**(1), 31–44, 2007.

[8] SMILE2004, Sound Material in Living Environment, Architectural Institute of Japan and Gihodo Shuppan Co., Ltd., 2004.