Estimation of fundamental frequency of reverberant speech by utilizing complex cepstrum analysis

Masashi Unoki and Toshihiro Hosorogiya
18 June 2007
IS-RR-2007-008

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, JAPAN
unoki@jaist.ac.jp, t-hosoro@jaist.ac.jp

ISSN 0918-7553

RESEARCH REPORT

Estimation of fundamental frequency of reverberant speech by utilizing complex cepstrum analysis

Masashi Unoki and Toshihiro Hosorogiya

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan
E-mail: {unoki, t-hosoro}@jaist.ac.jp

Abstract **This paper reports the comparative evaluations of twelve typical methods of estimating fundamental frequency (F_0) over huge speech-sound datasets in artificial reverberant environments. They involve several classic algorithms such as Cepstrum, AMDF, LPC, and modified autocorrelation algorithms. Other methods involve a few modern instantaneous amplitude- and/or frequency-based algorithms, such as TEMPO, IFHC, and PHIA. The comparative results revealed that the percentage correct rates and SNRs of the estimated F0s were reduced drastically as reverberation time increased. They also demonstrated that homomorphic (complex cepstrum) analysis and the concept of the source-filter model were relatively effective for estimating F_0 from reverberant speech. This paper thus proposes a new method of robustly and accurately F_0 estimating in reverberant environments, by utilizing the MTF concept and the source-filter model on the complex cepstrum analysis. The MTF concept is used in this method to eliminate dominant reverberant characteristics from observed reverberant speech. The source-filter model (liftering) is used to extract source information from the processed cepstrum. Finally, F_0s are estimated from them by using the comb-filtering method. Additive-comparative evaluation was carried out on the proposed method with other typical methods. The results demonstrated that it was better than the previously reported methods in terms of robustness and providing accurate F_0 estimates in reverberant environments.**

Keywords: Fundamental frequency (F_0), F_0 estimation, reverberant speech, complex cepstrum analysis, MTF concept, source-filter model

1. Introduction

The fundamental frequency (F_0) as well as the fundamental period (T_0) of speech can be utilized as significant features to represent the source information (glottal waveform or vocal-fold vibrations) 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, estimating the F_0 of target speech in real environments, which is the same as extracting the F_0 of noiseless speech, is a particularly important issue in these applications. This is because accurate F_0 information can be used to resolve serious problems that occur in realistic speech-signal processing.

It is well known that noise and reverberation smear significant features of speech so that the recognition rates of ASR systems are drastically reduced as the SNR of noise increases and/or reverberation time increases [1, 2, 3]. This is because accurately estimated F_0 can be used for spectrum normalization [4], noise reduction [5], feature extraction [6], speech emphasis [7, 8], and speech dereverberation [9] to improve the ability of ASR systems. Hence, robust and accurate estimates of F_0s from target speech in real environments is the ultimate goal in this research field.

Many studies on extracting or estimating the F_0 of target speech have been done in the literature on speech signal processing, and many methods have been proposed [10, 11, 12, 13] over the last half century. The traditional extraction/estimation methods can be divided into processing in the time and frequency domains, or both domains. Most of these have made use of the periodic features of speech in the time domain (zero-cross [14, 15], periodgram [16], peak-picking [14, 17], autocorrelation [14, 18], AMDF [19], and maximum likelihood [20, 21]) or harmonic features in the frequency domain (comb filtering [22, 23, 24, 25], autocorrelation [26, 27], sub-harmonic summation [28], and cepstrum [29, 34]).

The aim of all these methods has been to extract

the periodicity or harmonicity of source information from observed speech. However, this still seems to be incompletely resolved because three main issues remain, i.e., (1) observability: the observed speech is an emission sound passing through the mouth/nose so that it is impossible to directly observe glottal vibrations from it without eliminating the effects of the vocal tract, (2) flexibility and irregularity: glottal vibrations are not complete periodic signals and the range of variations in the periods is relatively wide, and (3) 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 (vocal tract), based on the source-filter model, from the observed speech sounds. For example, typical approaches based on this idea have been homomorphic analysis (cepstrum) methods [29, 30, 31, 32, 33, 34] and LPC-methods [35, 36, 37, 38, 39]. A few examples of inverse filtering methods are moving average with band-limitation [40], Lag-windowing [41], SIFT [42], and compensation by temporal continuity [43]. Center-clipping and band-limitation [44, 45], and multi-windowing [46] techniques have also been used in approaches based on the autocorrelation function.

A few approaches to precisely estimating the F_0 of target noiseless speech have been established (e.g., STRAIGHT-TEMPO [47] and YIN [48]) by comparing electro-glottal-graph (EGG) information. The stability of the instantaneous frequency of speech has also been used in the STRAIGHT-TEMPO method (referred to as "TEMPO" after this) to accurately estimate F_0s as significant features to resolve the first two issues. This method plays an important role in controlling "pitch" related features in STRAIGHT analysis/synthesis tools [49]. YIN has also been proposed that combines autocorrelation functions and AMDF to resolve these. It has been reported that both methods can be used to estimate the F_0 of target noiseless speech extremely precisely 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 (noisy and/or reverberant) environments. Hence, we need to investigate the last issue for realistic applications.

It is generally known that the method of estimating F_0 using periodic and/or harmonic features (e.g., autocorrelation functions and comb filtering) is relatively robust against background noise, but the estimated F_0 is not relatively accurate [12, 50, 51, 52]. It has also been reported that the comb-filtering-based method is more robust against background noise than the autocorrelation-based one [52, 53]. The cepstrum-based method is not as robust against background noise as either of these because it is composed of homomorphic analysis so that noise components are not clearly separated in the quefrency domain [52, 53].

The time-frequency representation of speech obtained by time-frequency analysis can also adequately represent the periodic/harmonic components of speech [54]. The instantaneous amplitude of speech signals has fine harmonic features that are robust against background noise so that comb-filtering of instantaneous amplitude has been proposed [59, 60] to construct a sound segregation model. The instantaneous frequency of speech has also been used to accurately estimate F_0s [55] but their stability as used in TEMPO is sensitive to noise. More robust methods using instantaneous amplitude and frequency have been proposed by using post-processing (dynamic programming) [56] and bandwidth equations related to instantaneous amplitude and frequency with harmonicity [50, 51, 57, 58]. Other robust techniques using instantaneous amplitude and frequency-related approaches have been proposed by using periodicity and harmonicity [52]. 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 noise conditions to estimate sufficiently accurate F_0s of target speech. Thus, methods using instantaneous amplitude and frequency 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 reverberant environments.

It can easily be predicted that no typical methods will work as well and their percentage correct rates for F_0s are reduced drastically as reverberation time increases. If our prediction is correct, the last issue has not yet been completely solved and needs to be considered in reverberant environments and in noisy reverberant environments. We evaluate traditional methods of estimating F_0 in terms of robustness and accuracy in reverberant environments in this paper to investigate this issue. We then propose a method of estimating F_0 from reverberant speech by taking the characteristics of reverberation into consideration.

This paper is organized as follows. Section 2 describes the mathematical setup and then defines the problem of estimating F_0 from reverberant speech. We evaluate most typical methods of estimating F_0 in reverberant environments in Section 3 and investigate what the best model is. Section 4 introduces complex cepstrum analysis and investigates what the significant features for robust estimates are. We then introduce the model concept (complex cepstrum analysis, the modulation transfer function (MTF) concept, and source-filter model (liftering)). We finally propose a method of estimating F_0 in reverberant environments. We evaluate our proposed method in Section 5 by com-

paring it with other methods using the same simulations. Section 6 gives our conclusions and perspectives regarding further work.

2. Mathematical setup

2.1 Signal representation and STFT

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

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

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 so that $\omega_k(t)$ can be expressed as $2\pi k F_0(t)$. Fundamental frequency, $F_0(t)$, is an instantaneous frequency so that this should be extracted from $x(t)$ using instantaneous cues.

The short-term Fourier transform (STFT) is usually used to analyze $x(t)$ in any given short term segment (windowing processing): [61]

$$\begin{aligned} X(\omega,\tau) &= \int x(t)w(t-\tau)\exp(-j\omega t)dt, & (2)\\ &= A(\omega,\tau)\exp(j\arg\phi(\omega,\tau)), & (3)\\ A(\omega,\tau) &= |X(\omega,\tau)|, & (4)\\ \phi(\omega,\tau) &= \arctan\left(\frac{\Im[X(\omega,\tau)]}{\Re[X(\omega,\tau)]}\right), & (5) \end{aligned}$$

where $w(t)$ is a window function and a short-term signal, $x(t,\tau)$, is defined as $w(t-\tau)x(t)$ for mathematical convenience. $A(\omega,\tau)$ is the amplitude spectrum and $\phi(\omega,\tau)$ is the phase spectrum of $X(\omega,\tau)$.

The task of extracting/estimating the fundamental frequency $F_0(t)$ in this formulation is, therefore, to estimate the F_0 in each short-term segment using the harmonicity of $X(\omega,\tau)$ or to estimate segmental $T_0 = 1/F_0$ by using the periodicity of $x(t,\tau)$. Thus, traditional methods based on waveform processing (e.g., zero-cross [14, 15], periodgram [16], peak-picking [14, 17], autocorrelation [14, 18], AMDF [19], maximum likelihood [20, 21], STFT-based processes, and sub-harmonic summation (SHS) [28]) estimate $F_0(t)$ from $x(t,\tau)$ or $X(\omega,\tau)$ by using periodicity or harmonicity. These are listed in the first two row in Table 1.

2.2 Source-filter model

The source-filter model is a well-known concept to separately represent glottal (source information) and vocal-tract (filter information) characteristics for speech production (or speech synthesis). Based on this concept, the observed clean speech signal $x(t)$ can be represented as

$$x(t) = e(t) * v_\tau(t), \qquad (6)$$

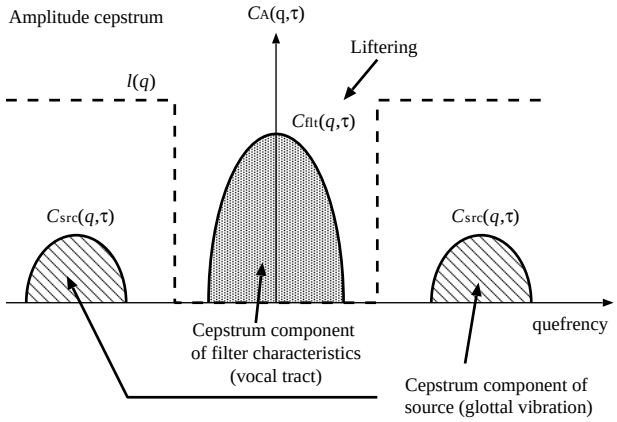

Fig. 1 Separated representations of source (glottal) and filter (vocal tract) characteristics in quefrency domain.

where $e(t)$ 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. "$*$" denotes convolution. Note that the emission effect has been omitted from this formulation. Thus, Eq. (2) can also be represented as

$$X(\omega,\tau) = S(\omega,\tau) \cdot V(\omega,\tau), \qquad (7)$$

where $S(\omega,\tau)$ is the STFT of $s(t,\tau) = w(t-\tau)e(t)$ and $V(\omega,\tau)$ is that of $v(t,\tau) = v_\tau(t)$. $V(\omega,\tau)$ represents filter characteristics so that the separation effect of $V(\omega,\tau)$ is usually used to estimate $F_0(t)$ from $X(\omega,\tau)$. Some traditional methods of estimation are inverse filtering $V^{-1}(\omega,\tau)$ [42], whitening of $X(\omega,\tau)$ by $|V(\omega,\tau)|$ (or lag windowing) [41], and subtraction on logarithmic processing $\log X(\omega,\tau) = \log S(\omega,\tau) + \log V(\omega,\tau)$ [44, 45]. These are listed in the second row in Table 1.

The linear prediction (LP) method is also one of the most powerful techniques of analyzing speech signals [35]. LP coefficients have filter characteristics (all-pole type) and LP residue has source information. The LP coefficients of $x(t,\tau)$ can thus be used as inverse filtering $V^{-1}(\omega,\tau)$ in the source-filter model [36, 37, 42]. LP residue can also be used as a short-term signal $s(t,\tau)$ [39]. Waveform processing and AMDF have also been incorporated [38]. These are listed in the third row in Table 1.

2.3 Cepstrum representation

Cepstrum is also a well-known method of homomorphic analysis. The complex cepstrum of $X(\omega,\tau)$ in Eq. (2) can be represented as

$$\begin{aligned} C(q,\tau) &= \mathcal{F}^{-1}\left[\log X(\omega,\tau)\right]\\ &= \mathcal{F}^{-1}\left[\log\left\{|X(\omega,\tau)|\exp(j\phi(\omega,\tau))\right\}\right]\\ &= \mathcal{F}^{-1}\left[\log A(\omega,\tau)\right] + \mathcal{F}^{-1}\left[j\phi(\omega,\tau)\right]\\ &= C_A(q,\tau) + C_\phi(q,\tau), \qquad (8) \end{aligned}$$

Table 1 Characteristics of typical methods of estimating F_0.

Algorithm	domain	periodicity	harmonicity	filter shape	Features
Waveform processing					
(1) zero-cross [14, 15]	time	o	x	x	$x(t,\tau)$
(2) peak detection [14, 17]	time	o	x	x	$x(t,\tau)$
(3) autocorrelation [18]	time	o	x	x	$x(t,\tau)$
(4) maximum likelihood [20, 21]	time	o	x	x	$x(t,\tau)$
(5) ACMWL [46]	time	o	x	x	$x(t,\tau)$
AMDF [19]	time	o	x	x	$x(t,\tau)$
YIN [48]	time	o	x	o	$s(t,\tau)$
STFT					
(1) auto-correlation [44, 45, 26]	freq.	x	o	x	$\log\lvert X(\omega,\tau)\rvert$
(2) Lag windowing [41]	freq.	x	o	o	$\lvert S(\omega,\tau)\rvert$
(3) Comb filtering method [22, 23, 25]	freq.	x	o	x	$\lvert S(\omega,\tau)\rvert$
SHS [28]	freq.	x	o	x	$\log\lvert X(\omega,\tau)\rvert$
LPC					
(1) Residue [39]	time	o	x	o	$s(t,\tau)$
(2) SIFT [42]	freq.	x	o	o	$\lvert S(\omega,\tau)\rvert$
Cepstrum					
(1) Noll's method [29, 31]	quef.	o	x	o	$C_A(q,\tau)$
(2) Clipstrum [32]	quef.	o	x	o	$C_A(q,\tau)$
(3) Improved cepstrum [40]	quef.	o	x	o	$C_A(q,\tau)$
(4) liftering method (this paper)	quef.	x	o	o	$C_S(\omega,\tau)\rvert$
F_0 filtering [64]	time	o	o	x	$s(t,\tau)$
IF-based method					Instant. freq. (IF)
(1) TEMPO [47]	freq.	x	x	x	Fixed point analysis
(2) IFHC [50, 51]	freq.	x	o	x	Harmonicity of IFs
(3) DASH [57, 58]	freq.	x	o	x	Degree of dominance
IA-based method					Instant. amp. (IA)
(1) Abe *et al.*'s method [56]	freq.	o	o	x	post-processing (DP)
(2) PHIA [52]	time/freq.	o	o	x	Dempster's law
Proposed method	time/freq./quef.	o	o	o	$s(t,\tau)$

where $C_A(q,\tau)$ is the amplitude cepstrum and $C_\phi(q,\tau)$ is the phase cepstrum of $C(q,\tau)$. q denotes quefrency (time domain). The complex cepstrum of $X(\omega,\tau)$ in Eq. (7) can also be represented as

$$\begin{aligned} C(q,\tau) &= \mathcal{F}^{-1}\left[\log S(\omega,\tau)\right] + \mathcal{F}^{-1}\left[\log V(\omega,\tau)\right] \\ &= C_{\rm src}(q,\tau) + C_{\rm flt}(q,\tau), \end{aligned} \tag{9}$$

where $C_{\rm src}(q,\tau)$ is the complex cepstrum of source $S(\omega,\tau)$ and $C_{\rm flt}(q,\tau)$ is that of filter $V(\omega,\tau)$.

The amplitude cepstrum, $C_A(q,\tau)$, is generally used in the traditional method so that $C_{A,\rm src}(q,\tau)$ and $C_{A,\rm flt}(q,\tau)$ are separately used for estimating $F_0(t)$ from $C_A(q,\tau)$. Figure 1 outlines the concept underlying the source-filter model in the quefrency domain. $C_{A,\rm flt}(q,\tau)$ represents the dominant spectrum envelope of $X(\omega,\tau)$ (lower Fourier component in quefrency domain) so that they are compactly located in the lower quefrency. In contrast, $C_{A,\rm src}(q,\tau)$ represents dominant fine structure of $X(\omega,\tau)$ so that they are compactly located in the higher quefrency domain. Therefore, the task of estimating F_0 with this concept is to find the dominant quefrency from $C_{A,\rm src}(q,\tau)$ or to detect periodicity or harmonicity from $C_{A,\rm src}(q,\tau)$ by eliminating $C_{A,\rm flt}(q,\tau)$ from $C_A(q,\tau)$. The last processing is referred to as "liftering". Typical approaches are Noll's original method [29, 31], his clipstrum method [32], and Kato and Miwa's improved method [40]. These are listed in the fourth row in Table 1.

2.4 Problem with estimating F_0

The task of estimating F_0 in reverberant environments is to extract $F_0(t)$ from reverberant speech signal $y(t)$ or respective STFT $Y(\omega,\tau)$:

$$y(t) = x(t) * h(t) = e(t) * v_\tau(t) * h(t), \tag{10}$$

$$\begin{aligned} Y(\omega,\tau) &= X(\omega,\tau)H(\omega,\tau) \\ &= S(\omega,\tau)V(\omega,\tau)H(\omega,\tau), \end{aligned} \tag{11}$$

where $h(t)$ is the impulse response and $H(\omega,\tau)$ is the STFT of $h(t)$ in room acoustics (reverberation). Note that, $H(\omega,\tau)$ is actually required to present all characteristics ($H(\omega) = H(\omega,\tau)$) by using long-term Fourier transform (LTFT) so that the length of analysis (at each τ) should be over the reverberation time.

The task of estimating F_0 in reverberant environments is thus to select periodicity and harmonicity from the convolved source signal, $e(t)$, while that in noisy environments is to select them from the noisy (additive) source signal, $e(t)$. If $h(t)$ is simplified echo or minimum phase impulse response, the cepstrum-based method can be used to adequately estimate F_0 from the reverberant speech signal, $y(t)$, because homomorphic analysis is a powerful tool for dealing with

simplified echos. Realistic impulse responses in room acoustics generally have non-minimum phase characteristics and we therefore predicted that estimating F_0 robustly and accurately would be more difficult than in noisy environments.

3. Evaluation of typical methods

3.1 Typical methods of estimating F_0

Many methods of estimating F_0 have been proposed in the literature on speech signal processing, as described in Section 1. The most comprehensive review remains that of Hess (1983) [11] and more recent reviews are those of Suzuki (1997), Hess (1992), and Cheveigné and Kawahara (2001) [10, 12, 13]. A few examples of recent approaches are instantaneous-amplitude [56, 59, 60], instantaneous-frequency [50, 51, 57, 58], fundamental wave-filtering [64], and wavelet methods [65], as well as auditory models [66, 67]. There are also comparative evaluations in Atake *et al.*'s (2000), Ishimoto *et al.*'s (2001, 2005), and Nakatani and Irino (2002, 2004) [12, 50, 51, 52, 13, 57, 58, 53].

We evaluated twelve typical methods to investigate how robust estimates of F_0 were in reverberant environments:

1. ACMWL (AutoCorrelation through Multiple Window-Length) [46]
2. AMDF (Averaged Magnitude Difference Function) [19]
3. STFT-ACorrLog (AutoCorrelation of Log-amplitude spectrum on STFT) [44, 45, 26]
4. STFT-ACorrLag (Lag-windowing of amplitude spectrum on STFT) [41]
5. STFT-Comb (Comb filtering of amplitude spectrum on STFT) [22, 23, 25]
6. SHS (Sub-Harmonic Summation) [28]
7. Cepstrum (Improved cepstrum) [29, 31]
8. LPC-residue (autocorrelation on LPC residue) [39]
9. VFWFF (Voice Fundamental Wave Filtering (Feed forward type)) [64]
10. TEMPO [47]
11. IFHC (Instantaneous Frequency of Harmonic Components) [50, 51]
12. PHIA (Periodicity/Harmonicity using Instantaneous Amplitude) [52]

All these methods are listed in Table 1. Although other methods have been proposed, we choose these twelve because they are commonly used in comparative evaluations and the others are just modifications or heavy revisions of them.

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.* [50, 51]. This dataset consisted of 30 short Japanese sentences uttered by 14 males and 14 females with voiced-unvoiced labels (total of 840 utterances, total duration of 40 min, sampling frequency of 16 kHz, and quantization of 16-bits).

The reverberant speech sentences we used were created by convolving the original signals, $x(t)$s, with the following reverberant impulse responses, $h(t)$s, as a function of the reverberation time.

$$h(t) = a \exp\left(\frac{-6.9t}{T_R}\right) n(t), \tag{12}$$

$$a = \left[1 \Big/ \int_0^T \exp\left(\frac{-13.8t}{T_R}\right) dt\right]^{1/2}, \tag{13}$$

where a is a constant gain factor as the normalized power of $h(t)$, T_R is reverberation time, and $n(t)$ is white noise. This is a formulation for the impulse response of artificial reverberation and has non-minimum phase components [62, 63]. 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.

Fine F_0 error and gross F_0 error have been used as measures for some comparative evaluations in noisy environments [12, 50, 52, 58], These have been concentrated into error analysis. Since we concentrated on evaluating robustness and the accuracy of F_0 estimates, we used two similar measures for evaluation but not the same measures. The first was the percent correct rate (expressed as %) and the second was SNR (in dB).

$$\text{Correct rate}_E = \frac{N_{F_{0,\text{Est}}}(E)}{N_{F_{0,\text{Ref}}}} \times 100, \tag{14}$$

$$\text{SNR} = 20 \log_{10} \frac{\int (F_{0,\text{Ref}}(t) - F_{0,\text{Est}}(t))^2 dt}{\int F_{0,\text{Ref}}(t)^2 dt}, \tag{15}$$

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

$$\frac{|F_{0,\text{Ref}}(t) - F_{0,\text{Est}}(t))|}{F_{0,\text{Ref}}(t)} \leq E, \qquad (\%)$$

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 estimated by TEMPO from the EGG 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 twelve methods from reverberant (or noiseless) speech signals. Two values for E (error margins of 5% and 10%) were used in the percent correct rate.

Since gross F_0 error is the ratio of the number of frames giving "incorrect" F_0 values to the total number of frames, the percent correct rate indicates approximately gross F_0 error. Since fine F_0 error is the normalized room mean square error between $F_{0,\mathrm{Ref}}(t)$ and $F_{0,\mathrm{Est}}(t)$, SNR indicates a similar measure in dB.

3.3 Results

Figure 2 plots the results of comparative evaluations for the twelve typical methods of estimating F_0 from reverberant speech as a function of the reverberation time. The left panels (a), (c), and (e) plot the results for the first six methods and the right panels (b), (d), and (f) plot them for the last six. The top panel plots the percent correct rates (expressed as percentages) for F_0 estimates within an error margin of 5% and the middle panel plots these within an error margin of 10%. The bottom panel plots the SNRs. The correct rates and SNRs of all 12 methods are drastically reduced as the reverberation time increases. The correct rates within the 5% error margin for all methods were less than 50% and the SNRs were less than about 15 dB, especially when reverberation time T_R was 2.0 s. Moreover, the correct rates within the 10% error margin as an approximate evaluation were also less than 70%. We hence concluded that none of these methods worked as well as robust and accurate F_0 estimates and they had drawbacks in estimating F_0 from reverberant speech.

However, we found a few clues in this evaluation for improving these methods. We can see from Fig. 2 that the cepstrum method is the most accurate excluding the clean condition ($T_R = 0.0$). Cepstrum analysis is homomorphic and this can deal with convolution processing as additive (subtractive) processing. Although the impulse responses we used in evaluations were not minimum-phase characteristics, the cepstrum method seemed to reduce the effect of reverberation for estimating F_0 since this can treat a direct sound and a reflected sound as the same signal. Therefore, the cepstrum method has the possibility of estimating F_0 from reverberant speech if it is not affected too much by reverberation. The comb-filtering method is slightly robust a reverberation as we can see from Figs. 2(c) and (e). Maximization of matched harmonicity may have the effect of tracking stationary fluctuations of harmonics that are not often affected by reverberation.

4. Proposed method

4.1 Complex cepstrum analysis

Let us overview the results in Subsection 3.3 by reconsidering the complex cepstrum representation of the reverberant speech $y(t)$. From Eqs. (9)-(11), the complex cepstrum of $y(t)$ can be represented as

$$\begin{aligned} C_Y(q,\tau) &= C_X(q,\tau) + C_H(q,\tau) \\ &= C_{\mathrm{src}}(q,\tau) + C_{\mathrm{flt}}(q,\tau) + C_H(q,\tau), \end{aligned} \tag{16}$$

where $C_H(q,\tau)$ is the complex cepstrum of the reverberant impulse response, $h(t)$. These cepstra can also be represented as all amplitude and phase cepstra (denoted by subscripts "A" and "ϕ").

The complex cepstrum analysis, on the other hand, is usually used to separate minimum and non-minimum (all-pass) phase characteristics. The complex cepstrum, $C(q,\tau)$, can also be separately represented as

$$\begin{aligned} C(q,\tau) &= C_{\min}(\omega,\tau) + C_{\mathrm{all}}(\omega,\tau) \\ &= C_{A,\min}(q,\tau) + C_{\phi,\min}(q,\tau) \\ &\quad + C_{A,\mathrm{all}}(q,\tau) + C_{\phi,\mathrm{all}}(q,\tau), \end{aligned} \tag{17}$$

where the subscripts "min" and "all" indicate minimum and non-minimum (all-pass) phase characteristics. Figure 3 is a schematic of the complex cepstrum. Here, as respective spectra can be represented as

$$\begin{aligned} X(\omega,\tau) &= X_{\min}(\omega,\tau) \cdot X_{\mathrm{all}}(\omega,\tau) \\ &= |X_{\min}(\omega,\tau)| \exp(j\phi_{\min}(\omega,\tau)) \\ &\quad \times |X_{\mathrm{all}}(\omega,\tau)| \exp(j\phi_{\mathrm{all}}(\omega,\tau)), \end{aligned} \tag{18}$$

the amplitude spectrum $|X_{\mathrm{all}}(\omega,\tau)| = 1$ and $C_{A,\mathrm{all}}(q,\tau) = 0$. Figure 4 plots the transform relations between short-term waveforms and the complex cepstrum via the complex spectrum.

Hence, a complete representation of $C_Y(q,\tau)$ can be separately represented as

$$\begin{aligned} &C_{Y,A,\min}(q,\tau) + C_{Y,\phi,\min}(q,\tau) + C_{Y,\phi,\mathrm{all}}(q,\tau) \\ &= C_{\mathrm{src,A,min}}(q,\tau) + C_{\mathrm{src},\phi,\min}(q,\tau) + C_{\mathrm{src},\phi,\mathrm{all}}(q,\tau) \\ &\quad + C_{\mathrm{flt,A,min}}(q,\tau) + C_{\mathrm{flt},\phi,\min}(q,\tau) + C_{\mathrm{flt},\phi,\mathrm{all}}(q,\tau) \\ &\quad + C_{H,A,\min}(q,\tau) + C_{H,\phi,\min}(q,\tau) + C_{H,\phi,\mathrm{all}}(q,\tau). \end{aligned} \tag{19}$$

Note that the amplitude cepstrum of all-pass phase characteristics have been omitted from this equation.

According to Eq. (16), 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 $C_H(q,\tau)$ (on LTFT), in which the length of analysis is over the reverberation time, is needed to accurately extract $C_{\mathrm{src}}(q,\tau)$.

We did a preliminary investigation into which component, $C_{H,\min}(q,\tau)$ or $C_{H,\mathrm{all}}(q,\tau)$, affected dealing with $C_{\mathrm{src}}(q,\tau)$ for estimating F_0, using Eq. (19). Figure 5 shows the process for estimating one of the reverberant speech signals (/Tokushima-To-Ieba-Awa-Odori-Ga-Yuumei-Desu/, female speaker, reverberation time T_R of 2.0 s) we used in the evaluations. Clean speech signals ($x(t)$ and reverberant $y(t)$) are

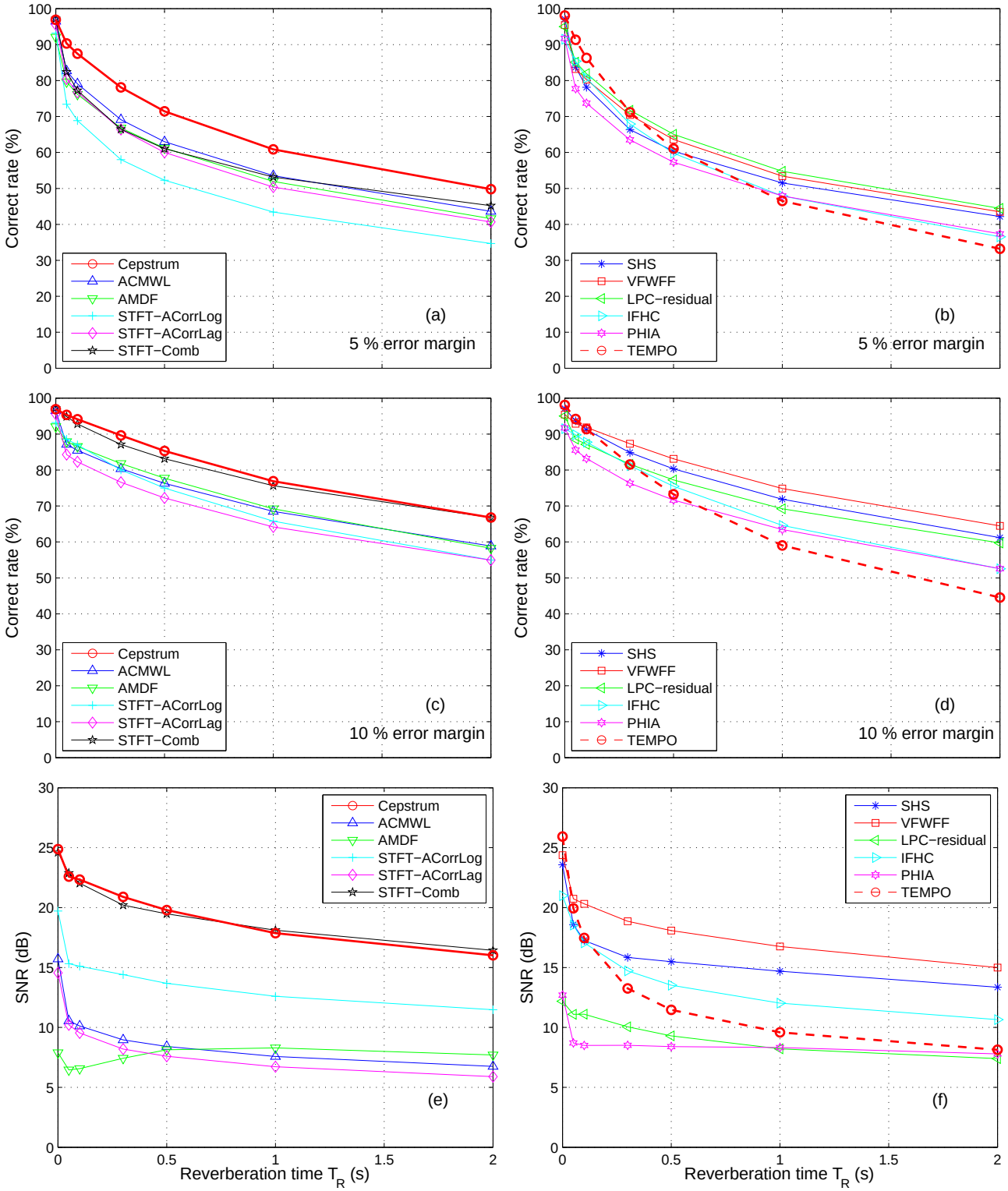

Fig. 2 Estimation results: (a)-(b) percent correct rate within error margin of 5% and (c)-(d) SNR (s: original, n: error between original and estimated F_0) of F_0 estimates from reverberant speech using twelve typical methods as function of reverberation time, T_R.

$x(t,\tau)$ (Periodic)	$=$	$x_{\min}(t,\tau)$ (Minimum-Phase Component)	$*$	$x_{\mathrm{all}}(t,\tau)$ (All-Pass Component)

(Time domain)

$\Downarrow \mathcal{F}$ $\Uparrow \mathcal{F}^{-1}$

$X(\omega,\tau)$ (Complex)	$=$	$X_{\min}(\omega,\tau)$ (Complex)	$\times$	$X_{\mathrm{all}}(\omega,\tau)$ (Complex)
$\|\|$		$\|\|$		$\|\|$
$\|X(\omega,\tau)\|$ (Real)	$=$	$\|X_{\min}(\omega,\tau)\|$ (Real)	$\times$	$\|X_{\mathrm{all}}(\omega,\tau)\|$ (Real)
$\times$		$\times$		$\times$
$e^{j\phi(\omega,\tau)}$ (Complex)	$=$	$e^{j\phi_{\min}(\omega,\tau)}$ (Complex)	$\times$	$e^{j\phi_{\mathrm{all}}(\omega,\tau)}$ (Complex)

(Frequency domain)

$\Downarrow \log$ $\Uparrow \exp$

$\log X(\omega,\tau)$ (Complex)	$=$	$\log X_{\min}(\omega,\tau)$ (Complex)	$+$	$\log X_{\mathrm{all}}(\omega,\tau)$ (Complex)
$\|\|$		$\|\|$		$\|\|$
$\log\|X(\omega,\tau)\|$ (Real)	$=$	$\log\|X_{\min}(\omega,\tau)\|$ (Real)	$+$	$\log\|X_{\mathrm{all}}(\omega,\tau)\|$ (Real)
$+$		$+$		$+$
$j\phi(\omega,\tau)$ (Imaginary)	$=$	$j\phi_{\min}(\omega,\tau)$ (Imaginary)	$+$	$j\phi_{\mathrm{all}}(\omega,\tau)$ (Imaginary)

(Frequency domain)

$\Downarrow \mathcal{F}^{-1}$ $\Uparrow \mathcal{F}$

$C(\omega,\tau)$ (Asymmetric)	$=$	$C_{\min}(\omega,\tau)$ (Asymmetric)	$+$	$C_{\mathrm{all}}(\omega,\tau)$ (Asymmetric)
$\|\|$		$\|\|$		$\|\|$
$C_A(\omega,\tau)$ (Even func.)	$=$	$C_{A,\min}(\omega,\tau)$ (Even func.)	$+$	$C_{A,\mathrm{all}}(\omega,\tau)$ (Even func.)
$+$		$+$		$+$
$C_\phi(\omega,\tau)$ (Odd func.)	$=$	$C_{\phi,\min}(\omega,\tau)$ (Odd func.)	$+$	$C_{\phi,\mathrm{all}}(\omega,\tau)$ (Odd func.)

(Quefrency (time) domain)

Fig. 3 Transform relations between waveform and complex cepstrum via complex spectrum. $\mathcal{F}$ means Fourier transform and $\mathcal{F}^{-1}$ means inverse Fourier transforms.

.

shown in Figs. 5(a) and (b). The reference F_0 ($F_{0,\mathrm{Ref}}(t)$ by TEMPO from the EGG signal) and the F_0 ($F_{0,\mathrm{Est}}(t)$) estimated by the cepstrum method from $y(t)$ are indicated in Fig. 5(c) by the dashed and solid lines. As can be seen, the estimated F_0 was not close to the reference. This method, however, can accurately estimate F_0 from $y(t)$ by eliminating the effect of $h(t)$ from $y(t)$ on the complex cepstrum in the long-term Fourier transform (LTFT), as plotted in Fig. 5(d). At the same time, two comparative F_0s were obtained as plotted in Figs. 5(e) and (f) by estimating F_0 from $y(t)$ by eliminating minimum phase or the all-pass phase component from $y(t)$.

The all-pass phase component of the reverberant impulse response $h(t)$ we used seems to have a dominant effect from these comparisons on robust and accurate F_0 estimates. Although the same comparisons for all the other stimuli are not presented in this paper, the same trends were observed. Hence, we concluded that eliminating the all-pass phase characteristics of $h(t)$ would enable effective estimates of F_0 from reverberant speech $y(t)$. In addition, the cepstrum method

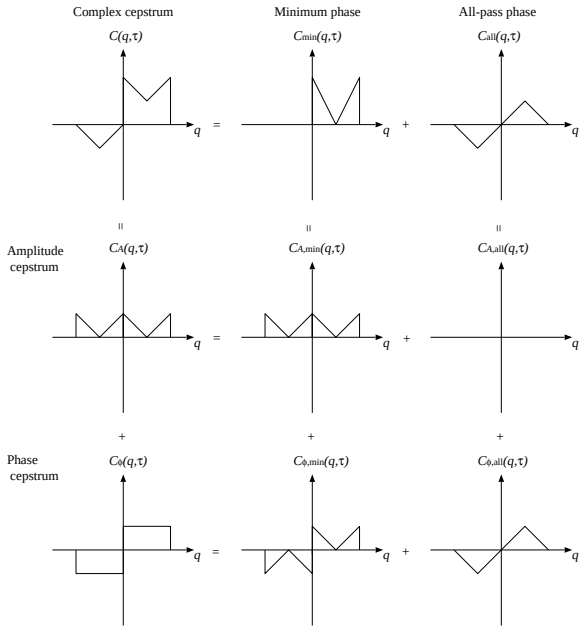

Fig. 4 Schematic of complex cepstrum relations: amplitude/phase cepstrum and minimum-phase/allpass-phase cepstrum.

with the all-pass component eliminated raises the possibility of achieving robust and accurate estimates of F_0 since we know homomorphic analysis can easily deal with minimum phase characteristics such as simplified echos.

4.2 Estimates of $h(t)$ based on MTF concept

The MTF concept was proposed by Houtgast and Steeneken [63] to account for the relation between the transfer function of frequency in an enclosure in terms of the envelopes of input and output signals ($x(t)$ and $y(t)$), and characteristics of the enclosure such as reverberation. This concept was introduced as a measure in room acoustics to assess what effect enclosure had on the intelligibility of speech [63]. The complex modulation transfer function, $\mathbf{m}(\omega)$, is defined as

$$\mathbf{m}(\omega) = \frac{\int_0^\infty h(t)^2 \exp(j\omega t)dt}{\int_0^\infty h(t)^2 dt}. \qquad (20)$$

This means the Fourier transform of the squared impulse response is divided by its total energy.

When reverberant impulse response $h(t)$ as defined in Eq. (12) is substituted into the equation above, the MTF, $m(\omega)$, can be obtained as

$$m(\omega) = |\mathbf{m}(\omega)| = \left[1 + \left(\omega \frac{T_R}{13.8}\right)^2\right]^{-1/2}. \qquad (21)$$

This means that $C_{H,A}(q,\tau)$ can be obtained from $\log|m(\omega)|$ with the power factor on the LTFT. Therefore, if T_R can be known without measuring $h(t)$, am-

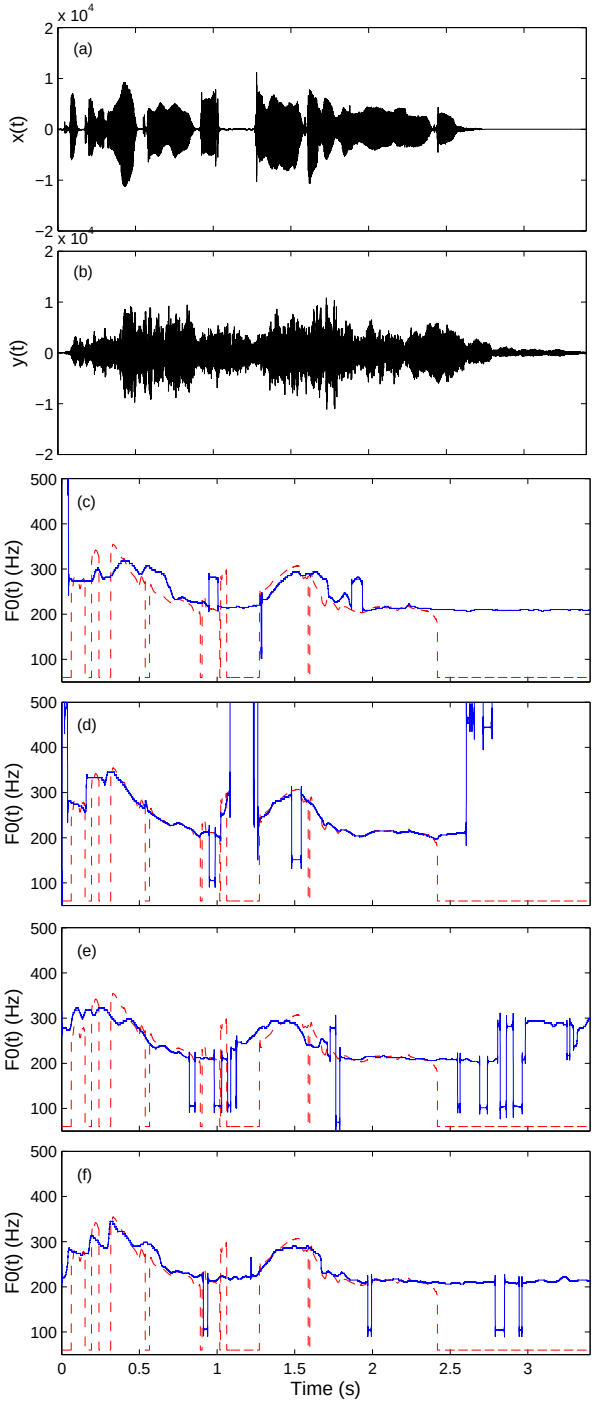

Fig. 5 Example: (a) original speech $x(t)$, (b) reverberant speech $y(t)$ (reverberation time of 2.0 s), (c) reference F_0 using TEMPO from EGG of $x(t)$ indicated by dashed-line and the estimated F_0 using cepstrum method from $y(t)$ indicated by solid line, (d) estimated F_0 from the dereverbed $y(t)$ using $h^{-1}(t)$, (e) $\hat{F}_0$ from $y(t)$ eliminated by minimum phase characteristics, and (f) $\hat{F}_0$ from $y(t)$ eliminated by all-pass phase characteristics.

plitude cepstrum $C_{H,A}(q,\tau)$ can be predicted by utilizing the MTF concept. The temporal envelope of the reverberant impulse response, $a\exp(-6.9t/T_R)$, can be also predicted with them.

MTF-based speech dereverberation methods, on the other hand, have been proposed by the present authors [68, 69]. A method of obtaining T_R estimates from reverberant speech $y(t)$ have also been proposed for blind-speech dereverberation. Fortunately, the method of obtaining T_R estimates can be applied to predicting $C_{H,A}(q,\tau)$ as well as the temporal envelope of $h(t)$ by using:

$$\hat{T}_R = \max\left(\underset{T_{R,\min}\le T_R \le T_{R,\max}}{\arg\min} \int_0^T \left| \min\left(\hat{e}_{x,T_R}(t)^2, 0 \right) \right| dt \right), \tag{22}$$

where T is the signal duration and $\hat{e}_{x,T_R}(t)^2$ is the set of candidates for the restored power envelope via inverse MTF [68] as a function of T_R. Note that the operation of "$\max(\arg\min\{\cdot\})$" means the maximum argument of T_R needs to be determined from a timing point where the negative area of $\hat{e}_{x,T_R}(t)^2$ approximately equals zero or a particular minimum area. Here, $T_{R,\min}$ and $T_{R,\max}$ are the lower and upper limited regions of T_R [68, 69].

According to Eqs. (12) and (22), $h(t)$ can be estimated by utilizing $\hat{a}\exp(-6.9t/\hat{T}_R)$ with simulated white noise $\hat{n}(t)$. This is referred to as $\hat{h}(t)$. In this case, long-term $C_H(q,\tau)$ can be obtained from $\hat{h}(t)$. Although this does not correspond to the original $h(t)$ we used in the evaluation, long-term amplitude cepstrum $C_{H,A}(q,\tau)$ can only be matched to the original. Although it is difficult to obtain a complete value with regard to phase cepstrum $C_{H,\phi}(q,\tau)$, long-term $C_{H,\phi,\mathrm{all}}(q,\tau)$ can be estimated from them by using Eqs. (17) and (19). As shown in Sec. 4.1, using estimated $C_{H,\phi,\mathrm{all}}(q,\tau)$ from $\hat{h}(t)$ to eliminate the all-pass phase component from reverberant speech $y(t)$ on the LTFT basis should be done to estimating F_0. Although the estimated $C_{H,A,\min}(q,\tau)$ can also be canceled out in Eq. (19) on LTFT, the elimination of minimum-phase characteristics in Eq. (19) on LTFT is not as effective for eliminating all-pass phase characteristics so that this is not used in this paper. Short-term $C_{H,A,\min}(q,\tau)$ and $C_{H,\phi,\min}(q,\tau)$ to be canceled out in Eq. (19) on STFT will be considered in the next section.

4.3 Liftering on complex cepstrum

$C_{H,\phi,\mathrm{all}}(q,\tau)$ is canceled out in Eq. (19) on LTFT as explained in the previous section, so that the remaining terms are $C_{\mathrm{flt}}(q,\tau)$ and $C_{H,\min}(q,\tau)$ to extract $C_{\mathrm{src}}(q,\tau)$. Complex cepstrum analysis and the source-filter model are used to cancel the remaining terms in Eq. (19) on STFT to take the best advantage of homomorphic processing.

There is a Hilbert transform relationship between $C_{A,\min}(q,\tau)$ and $C_{\phi,\min}(q,\tau)$, and $C_{H,\phi,\min}(q,\tau)$ has

the same characteristics in the positive quefrency domain based on the minimum phase characteristics. However, short-term $C_{H,A,\min}(q,\tau)$ and $C_{H,\phi,\min}(q,\tau)$ are not the same as the long-term versions when the length of STFT analysis is shorter than the reverberation time. However, amplitude cepstrum $C_{H,\min}(q,\tau)$ in the lower quefrency parts is generally larger than those in the higher parts and this attenuates exponentially as the quefrency increases. Therefore, the minimum phase characteristics, $C_{H,\min}(q,\tau)$, are assumed to concentrate on lower quefrency parts.

The cepstrum components of the source characteristics are separately concentrated on the higher quefrency parts and those of filter are separately concentrated on the lower based on the advantage of the source-filter model, as shown in Fig. 1. Therefore, if a component on the low quefrency part can only be removed by liftering, the filter characteristics as well as the dominant components of the minimum phase characteristics of reverberation can be canceled out in Eq. (19). Thus, the following lifter, $l(q)$, is used in this paper to cancel them out in Eq. (19).

$$l(q) = \begin{cases} 0, & q \le q_{\text{lif}} \\ 1, & q > q_{\text{lif}} \end{cases} \tag{23}$$

where $q_{\text{lif}} = 1.25$ ms. This means the upper limited estimated F_0 is 800 Hz.

4.4 Proposed method of estimating F_0

The algorithm for estimating F_0 based on complex cepstrum analysis, the MTF concept, and the source-filter model are explained in Fig. 6. This method is composed of three main processes: (1) estimating the MTF-based reverberation impulse responses and eliminating them from reverberant speech, (2) extracting $X_{\text{src}}(\omega,\tau)$ from the processed reverberant speech by using liftering on the complex cepstrum based on the source-filter model, and (3) estimating F_0 from them by using a final decision block.

Comb filtering was employed in the final two blocks in Fig. 6. As these are commonly used in classical methods of estimation, such as comb filtering and autocorrelation functions, they can be replaced by the autocorrelation function. In addition, since the proposed method treats a complex cepstrum, the restored short-term waveform $s(t,\tau)$ from $C_{\text{src}}(q,\tau)$ can be used to estimate F_0 with the autocorrelation function and/or AMDF. The aim of this paper was to propose a model concept for robustly estimating F_0 in reverberant environments. Therefore, these kinds of considerations with regard to the modification of processing are beyond the scope of this paper.

5. Evaluation of the proposed method

5.1 Method

We evaluated the proposed method with (labeled "Proposed(Est)") and without (labeled "Proposed(Org)") T_R estimates by using the same procedure and sound dataset described in Section 3. With and without comparisons of the proposed method were done to find how accurate the T_R estimates were. We compared them with TEMPO, the cepstrum method, and a modified complex cepstrum method based on the source-filter model (labeled "SrcFlt"). The SrcFlt method was used to find how effectively $C_{H,\phi,\text{all}}(q,\tau)$ was eliminated on the LTFT with the proposed method.

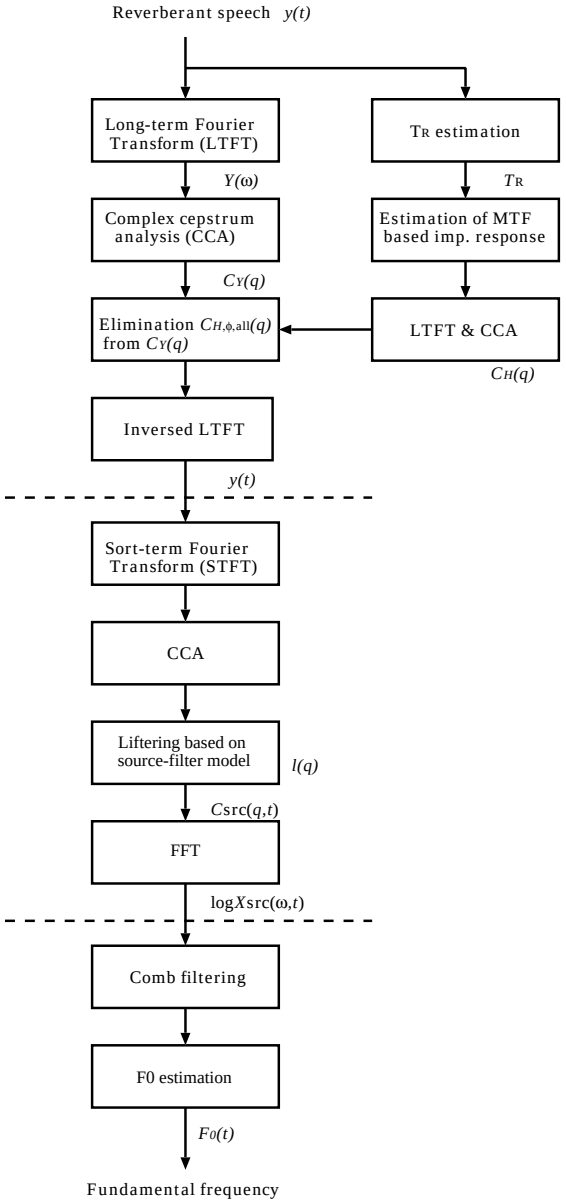

Fig. 6 Algorithm for proposed method.

5.2 Results and discussion

Figure 7 plots the results for the comparative evaluations. The correct rates within error margins of 5% and 10% for the proposed and the other methods are

plotted in Figs. 7(a) and (b). Their SNRs are plotted in Fig. 7(c). The results for the cepstrum method indicate the baselines in the evaluations while those for TEMPO (dashed-line) indicates the lower limits.

Although the overall accuracy of F_0 estimates tended to be reduced as reverberation time increased, about a 10% improvement in the correct rates and about a 5 dB improvement in the SNR 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 the correct rate of 60% within an error margin of 5%, the correct rate of 75% within an error margin 10%, and the SNR of 17 dB at $T_R = 2.0$ s, were achieved the method we propose, we concluded that MTF-based impulse responses can be precisely estimated by utilizing T_R estimates. For example, the results for extracting F_0 at $T_R = 2.0$ with the proposed method with and without T_R estimates from the same reverberant speech (Fig. 5(b)) are plotted in Figs. 7(d) and (e).

The SrcFlt method results indicate a small improvement (about 3% in the correct rate) to that with the cepstrum method. In contrast, there were about 7% and 5 dB improvments in the percent correct rate and in SNR 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.

6. Conclusion

We evaluated the robustness and accuracy of twelve typical methods of estimating F_0 (i.e., classic ACMWL, AMDF, STFT-based, cepstrum, LPC, and SHS algorithms, and modern IFHC, PHIA, and TEMPO algorithms) in artificial reverberant environments using huge speech datasets. The results revealed that none of these methods could accurately estimate F_0 in reverberant environments and that their accuracies drastically decreased as reverberation time increased. The results also demonstrated that the best method was cepstrum-based and that the worst was the instantaneous frequency-based model. We found that periodicity and/or harmonicity on the complex cepstrum were effective for estimating F_0 in reverberant environments.

We proposed a robust and accurate method of estimating F_0 that was based on the source-filter model concept and the MTF concept in complex cepstrum analysis. This method included (1) eliminating the dominant reverberation effect from observed speech by estimating MTF-based reverberant impulse responses and (2) extracting source information from them by subtracting the remaining cepstrum related to filter characteristics and the remaining reverberation through liftering. We demonstrated that our new method is robust against reverberation and can accurately estimate F_0 from observed reverberant speech, using the same comparative evaluations.

Additional improvements may be possible by modifying the F_0 determination block. Further evaluations using real reverberant impulse responses in room acoustics are required for real applications, but this is beyond the scope of this paper.

7. Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (No. 18680017). This work is also partially supported by SCOPE (071705001) of the Ministry of Internal Affairs and Communications (MIC), Japan. The authors would also like to thank Prof. M. Akagi and Dr. J. Li of JAIST for their helpful comments.

References

[1] S. Furui and M. M. Sondhi, Advances in Speech Signal Processing, New York Marcel Dekker, Inc., 1991.

[2] T. Takiguchi, S. Nakamura, and K. Shikano, "Hands-Free Speech Recognition by HMM composition in Noisy Reverberant Environments," IEICE Trans. D-II, Vol. J79-D-II, No. 2, pp. 2047–2053, Dec. 1997 (in Japanese).

[3] S. Nakagawa, "A Survey on Automatic Speech Recognition," IEICE Trans. D-II, Vol. J83-D-II, No. 2, pp. 433–457, Feb. 2000.

[4] H. Singer and S. Sagayama, "Pitch dependent phone modeling for HMM based on speech recognition," Proc. ICASSP92, Vol. 1, pp. 273–276, San Flancisco, CA, March 1992.

[5] J. C. Junqua, and J. P. Haton, ROBUSTNESS IN AUTOMATIC SPEECH RECOGNITION, – fundamentals and applications –, Kluwer Academic Publishers, Boston, 1996

[6] W. J. Hess, "A pitch-synchronous digital feature extraction system for phonemic recognition of speech," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 1, Feb. 1976.

[7] H. Hermansky, N. Morgan, and H. G. Hirsch. "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," ICASSP'93, 83–86, Mineapolic, April 1993.

[8] H. Hermansky and N. Morgen. "RASTA Processing of speech," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 4, pp. 578–589, Oct. 1994.

[9] X. Lu, M. Unoki, and M. Akagi, "A robust feature extraction based on the MTF concept for speech recognition in reverberant environment," Proc. ICSLP2006, pp. 2546–2549, Pittsburgh, USA, Sept. 2006.

[10] H. Suzuki, "A story of old-and news of pitch extraction in speech technology," J. Acoust. Soc. Jpn. Vol.56, No. 2, pp. 121–128, Feb. 2000.

[11] W. J. Hess, "Pitch Determination of Speech Signals," Springer-Verlag, New York, 1983.

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

[13] A. de Cheveigné and H. Kawahara, "Comparative evaluation of F0 estimation algorithms," Proc. Eurospeech2001, pp. 2451–2454, Scandinavia, Sept. 2001.

[14] B. Gold and L. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," J. Acoust. Soc. Am., Vol. 46, No. 2, pp. 442-448, Aug. 1969.

[15] N. C. Geckinli and D. Yavuz, "Algorithm for pitch extraction using zero-crossing interval sequence," IEEE Trans. Acoustics, Speech, and Signal processing, Vol. ASSP-25, No. 6, pp. 559–564, Dec. 1977.

[16] M. R. Schroeder, "Period histogram and product spectrum: new methods for fundamental frequency measurement," J. Acoust. Soc. Am., Vol. 43, No. 4, pp. 829-834, April 1968.

[17] D. M. Howard, "Peak-picking fundamental period estimation for hearing prostheses," J. Acoust. Soc. Am., Vol. 86, No. 3, pp. 902-910, Sept. 1989.

[18] M. M. Sondhi, "New methods of pitch extraction," IEEE Trans. Audio and Electroacoustics, Vol. AU-16, No. 2, pp. 262–266, June 1968.

[19] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-22, No. 5, pp. 353–361, Oct. 1974.

[20] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," IEEE Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-24, No. 5, pp. 418–423, Oct. 1976.

[21] D. H. Friedman, "Multidimensional Pseudo-Maximum-Likelihood Pitch Estimation," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-26, No. 3, pp. 185–196, June, 1978.

[22] K. Nishi and S. Ando, "An optimal comb filter for time-varying harmonics extraction," IEICE Trans. Fundamentals, Vol. E81-A, No. 8, pp. 1622–1627, Aug. 1998.

[23] K. Nishi and S. Ando, "Uniform-Q comb filter and its time/frequency characteristics – filter architecture for fluctuation error –" IEICE A, Vol. J81-A, No. 2, pp. 152–160, Feb. 2000 (in Japanese).

[24] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am., Vol. 93, No. 6, pp. 3271–3290, June 1993.

[25] T. Miwa, Y. Tadokoro, and T. Saito, "The pitch estimation of different musical instruments sounds using comb filters for transcription," IEICE, Trans. D-II, vol. J81-D-II, no. 9, pp. 1965–1974, Sept. 1998 (in Japanese).

[26] T. Shimamura and H. Kobayashi, "Weighted autocorrelation pitch extraction of noisy speech," IEEE Trans. Speech and Audio Processing, Vol. 9, No. 7, pp. 727–730, Oct. 2001.

[27] T. Shimamura and H. Takagi, "Fundamental frequency extraction method based on the p-th power of amplitude spectrum with band limitation," IEICE Trans. Fundamentals, Vol. J86-A, No. 11, pp. 1097–1107, Nov. 2003.

[28] D. J. Hermes, "Measurement of pitch by subharmonic summation," J. Acoust. Soc. Am., Vol. 83, No. 1, pp. 257–264, Jan. 1988.

[29] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," J. Acoust. Soc. Am., Vol. 36, No. 2, pp. 226–302, Feb. 1964.

[30] M. A. Poletti, "The Homomorphic Analysis Signal," IEEE Trans. Signal Processing, Vol. 45, No. 8, pp. 1943–1953, Aug. 1997.

[31] A. M. Noll, "Ceptrum pitch determination," J. Acoust. Soc. Am., Vol. 41, No. 2, pp. 293–309, Aug. 1966.

[32] A. M. Noll, "Clipstrum pitch determination," J. Acoust. Soc. Am., Vol. 44, No. 6, pp. 1585–1591, July 1968.

[33] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," IEEE Trans. Audio, Electroacoust., Vol. AU-16, No. 2, pp. 221–226, June 1968.

[34] A. V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," J. Acoust. Soc. Am., Vol. 45, No. 2, pp. 458–465, June 1969.

[35] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 3, pp. 247–254, June 1979.

[36] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and A. McGonegal. "A comparative study of several pitch detection algorithms," IEEE Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-24, pp. 399–413, 1976.

[37] J. D. Markel and A. H. Gray, "A linear prediction vocoder simulation based upon the autocorrelation method," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-22, No. 2, pp. 124–134, April 1974.

[38] C. K. Un, and S. C. Yang, "A pitch extraction algorithm based on LPC inverse filtering and AMDF," IEEE Trans. Acoust., Speech, Signal Process. Vol. ASSP-25, No. 6, pp. 565–572, Dec. 1977.

[39] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for indentification of closed glottis interval," IEEE Trans. Acoustics, Speech, Signal Processing, Vol. ASSP-27, No. 4, pp. 309–319, Aug. 1979.

[40] S. Kato, and J. Miwa, "Pitch detection using moving average and band-limitation in cepstrum method and its application," Tech. Rep. of IEICE, SP94-95, Feb. 1995.

[41] H. Singer, and S. Sagayama, "Pitch dependent phone modeling for HMM-based speech recognition," J. Acoust. Soc. Jpn. (E), Vol 15, No. 2, pp. 77–86, March 1994.

[42] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio, Vol. AU-20, No. 5, pp. 367–377, Dec. 1972.

[43] K. Yanagisawa, K. Tanaka, and I. Yamaura, "A detection method of fundamental period using time continuous properties of spectrum envelope," IEICE Trans. D-II, Vol. J83-D-II, No. 11, pp. 2087–2098, Nov. 2000 (in Japanese).

[44] N. Kunieda, T. Shimamura, and J. Suzuki, "Pitch extraction by using autocorrelation function on the log spectrum," IEICE Trans. A, Vol. J80-A, No. 3, pp. 435–443, March 1997 (in Japanese).

[45] H. Kobayashi and T. Shimamura, "An extraction method of fundamental frequency using clipping and band limitation on log spectrum," IEICE Trans. A, Vol. J82-A, No.

7, pp. 1115–1122, July 1999 (in Japanese).

[46] T. Takagi, N. Seiyama, and E. Miyasaka, "A Method for pitch extraction of speech signal using autocorrelation functions through multiple window-length," IEICE Trans. A, Vol. J80-A, No. 9, pp. 1341–1350, Sept. 1997 (in Japanese).

[47] H. Kawahara, H. Katayose, A. de Cheveigné, R. D. Patterson, "Fixed Point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," Proc.Eurospeech99, No. 6, pp. 2781–2784, Budapest, Hungary, Sept. 1999.

[48] A. de Cheveigné, H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., Vol. 111, No. 4, pp. 1917–1930, April 2002.

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

[50] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," Proc of ICSLP2000, Vol. 2, pp. 907–910, Beijing, China, Oct. 2000.

[51] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, K. Shikano, "Robust estimation of fundamental frequency using instantaneous frequencies of harmonic components," IEICE vol. J83-D-II, No. 11, pp. 2077–2086, Nov. 2000 (in Japanese).

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

[53] Y. Ishimoto, M. Unoki, and M. Akagi, "Fundamental frequency estimation for noisy speech based on instantaneous amplitude and frequency," JAIST Research Report, IS-RR-2005-006, March 2005.

[54] L. Cohen, Time-frequency analysis. Prentice hall PTR, New Jersey. 1995.

[55] J. C. Brown and M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the Fourier transform," J. Acoust. Soc. Am., Vol. 92, No. 2, pp. 662–667, Aug. 1993.

[56] T. Abe, T. Kobayashi, and S. Imai, "Pitch estimation based on instantaneous frequency in noisy environments," IEICE D-II, Vol. J79-D-II, No. 11, pp. 1771–1781, Nov. 1996 (in Japanese).

[57] T. Nakatani and T. Irino, "Robust fundamental frequency estimation against background noise and spectral distortion," Proc. ICSLP2002, pp. 1733–1736, Denver, Colorado, USA. Sept. 2002.

[58] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," J. Acoust. Soc. Am. Vol. 116, No. 6, pp. 3690–3700, Dec. 2004.

[59] M. Unoki and M. Akagi, "A method of extracting the harmonic tone from noisy signal based on auditory scene analysis," IEICE A, Vol. J82-A, No. 10, pp. 1497–1507, Oct. 1999 (in Japanese).

[60] M. Unoki and M. Akagi, "A Method of Signal Extraction from Noisy Signal based on Auditory Scene Analysis," Speech Communication, Vol. 27, No. 3, pp. 261–279, April 1999

[61] P. P. Vaidyanathan, "Multirate systems and Filter Banks," Prentice-Hall, New Jersey, 1993.

[62] H. Kuttruff, *Room Acoustics*, Taylor & Francis, fourth edition, London, 2000.

[63] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica, Vol. 28, pp. 66-73, (1973).

[64] H. Ohmura and K. Tanaka, "Fine pitch contour extraction by voice fundamental wave filtering method," J. Acoust. Soc. Jpn, Vol. 51, No. 7, pp. 509–518, July 1995 (in Japanese).

[65] A. Sasou, and S. Nakamura, "A pitch extraction method using wavelet transform," IEICE A, Vol. J80-A, No. 11, pp. 1848–1856, Nov. 1997 (in Japanese).

[66] L. M. Van Immerseel, and J. P. Martens, "Pitch and voiced/unvoiced determination with an auditory model," J. Acoust. Soc. Am., Vol. 91, No. 6, pp. 3511–3526, June 1992.

[67] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," J. Acoust. Soc. Am., Vol. 71, No. 3, pp. 679–688, March 1982.

[68] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," Acoustical Science and Technology, Vol. 25, No. 4, pp. 232–242, April 2004.

[69] M. Unoki, K. Sakata, M. Furukawa, and M. Akagi, "A speech dereverberation method based on the MTF concept in power envelope," Acoustical Science and Technology, Vol. 25, No. 4, pp. 243–254, April 2004.

Masashi Unoki was born in Akita Pref., Japan, in 1969. He received his M.S. and Ph.D. (Information Science) from the Japan Advanced Institute of Science and Technology (JAIST) in 1996 and 1999. His main research interests are auditory-motivated signal processing and the modeling of auditory systems. He was a JSPS research fellow from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories as a visiting researcher during 1999–2000, and from 2000 to 2001 he was a visiting research associate at CNBH in the Dept. of Physiology at the University of Cambridge. He has been on the faculty of the School of Information Science at JAIST since 2001 and he is now an associate professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electrical and Electronic Engineering (IEEE), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize for an Outstanding Paper from the ASJ in 1999 and the Yamashita Taro Prize for Young Researcher from the Yamashita Taro Research Foundation in 2005.

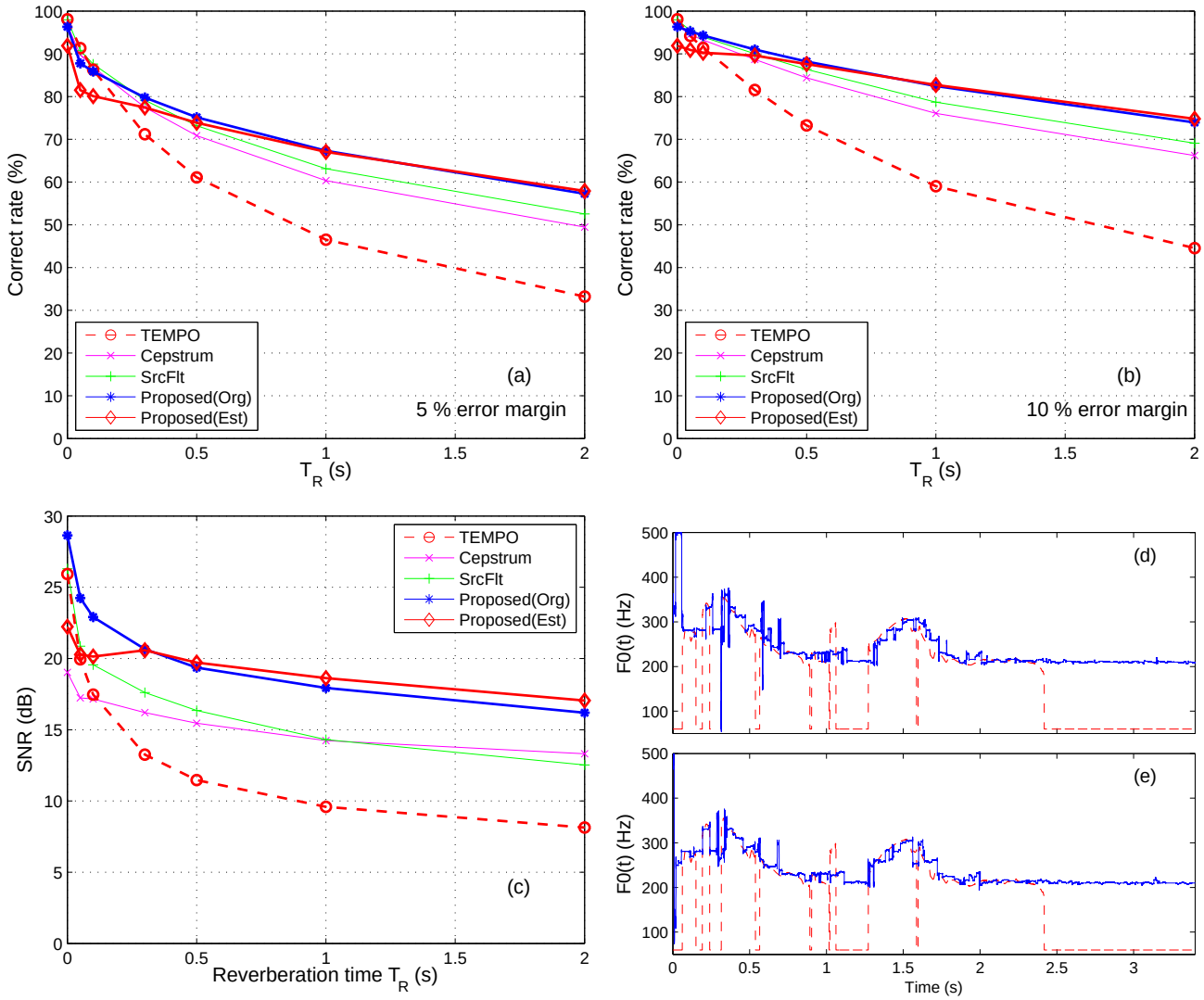

Fig. 7 Evaluation results: (a) percent correct rate within error margin of 5%, (b) percent correct rate within error margin of 10%, (c) SNR of F_0 estimation from reverberant speech using proposed method, and examples of extracted F_0 using proposed model (d) without T_R estimation and (e) with T_R estimation.

Toshihiro Hosorogiya was born in Ishikawa Pref., Japan in 1980. He received his B.E. from Nagoya University in 2005, and his M.S. from the Japan Advanced Institute of Science and Technology in 2007. He is a member of the Research Institute of Signal Processing (RISP) and the Acoustical Society of Japan (ASJ).