PAPER

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 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 STRAIGHT-TEMPO, IFHC, and PHIA. The comparative results revealed that the percentage of 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 estimating F_0s in reverberant environments, by utilizing the modulation transfer function (MTF) concept and the source-filter model in 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 new approach with other typical methods. The results demonstrated that it was better than the previously reported techniques in terms of robustness and providing accurate F_0 estimates in reverberant environments.**

Keywords: 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 sounds in various speech-signal processes. These are, for example, in speech analysis/synthesis systems, automatic speech recognition (ASR) systems, and speech emphasis methods. In particular, robust and accurate F_0 can generally be used as a powerful cue to reduce the noise component in noisy speech and/or to remove the reverberation effect in reverberant speech. Therefore, robustly and accurately 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.

Many studies on extracting or estimating the F_0 of target speech have been done in the literature on speech-signal processing, and numerous methods have been proposed [1, 2, 3] 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 [4, 5], periodgram [6], peak-picking [4, 7], autocorrelation [4, 8], the amplitude magnitude difference function (AMDF) [9], and maximum likelihood [10]), or harmonic features in the frequency domain (comb filtering [11, 12], autocorrelation [13], sub-harmonic summation (SHS) [14], and cepstrum [15]).

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 emitted 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 the 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 methods of analysis [15, 16] and linear prediction (LP)-based methods [17, 18, 19]. A few examples of inverse filtering methods are lag-windowing [20] and simplified inverse filter tracking (SIFT) [21]. Center-clipping, band-limitation [22, 23], and multi-windowing [24] techniques have also been used in approaches based on the autocorrelation function.

A few approaches to precisely estimate the F_0 of target noiseless speech have been established (e.g., STRAIGHT-TEMPO [25] and YIN [26]) 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 [27]. YIN has also been proposed which 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 have been resolved. However, it has not yet been clarified whether these methods can precisely estimate F_0 in real 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 [2, 28, 29]. It has also been reported that the comb-filtering-based method is more robust against background noise than the autocorrelation-based approach [29]. 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 [29].

The time-frequency representation of speech obtained by time-frequency analysis can also adequately represent the periodic/harmonic components of speech. The instantaneous amplitude (IA) of speech signals has fine harmonic features that are robust against background noise so that the comb-filtering of instantaneous amplitude has been proposed [30] to construct a sound-segregation model. 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 instantaneous frequency have been proposed by using bandwidth equations related to instantaneous amplitude and frequency with harmonicity [28, 31]. Other robust techniques using instantaneous amplitude and frequency-related approaches have been proposed by using periodicity and harmonicity [29]. 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 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 will drastically be reduced 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 evaluated traditional methods of estimating F_0 in terms of robustness and accuracy in reverberant environments to investigate this issue and discuss them in this paper. We then propose a method of estimating F_0 from reverberant speech without measuring impulse response in room acoustics (i.e., blind method of estimating F_0) 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 discuss our evaluations of most typical methods of estimating F_0 in reverberant environments in Section 3 and investigations into 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). We finally propose a method of estimating F_0 in reverberant environments. We discuss our evaluations of the proposed method in Section 5 by comparing it with

other methods using the same simulations. Section 6 gives our conclusions and perspective regarding future work.

2. Mathematical setup

2.1 Signal representation and STFT

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

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

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 [32] to analyze $x(t)$ in any given short-term segment (windowing processing):

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

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 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 [4, 5], periodgrams [6], peak-picking [4, 7], autocorrelation [4, 8], AMDF [9], maximum likelihood [10], STFT-based processes, and sub-harmonic summation (SHS) [14]) estimate $F_0(t)$ from $x(t,\tau)$ or $X(\omega,\tau)$ by using periodicity or harmonicity.

2.2 Source-filter model

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

$$x(t) = e(t) * v_\tau(t) \tag{6}$$

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

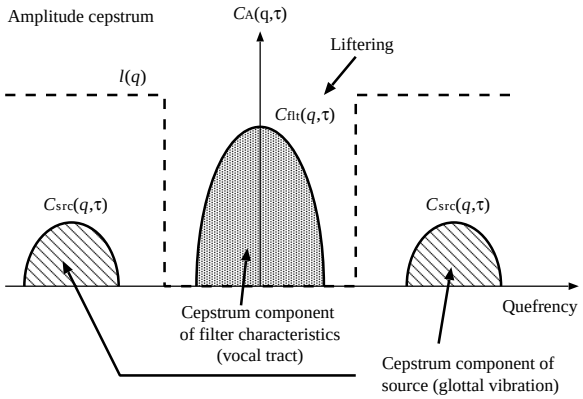

Fig. 1 Separated representations of source and filter characteristics in quefrency domain

$$X(\omega,\tau) = S(\omega,\tau)\cdot V(\omega,\tau) \tag{7}$$

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 $V^{-1}(\omega,\tau)$ [21], whitening of $X(\omega,\tau)$ using $|V(\omega,\tau)|$ (or lag windowing) [20], and subtraction on logarithmic processing $\log X(\omega,\tau) = \log S(\omega,\tau) + \log V(\omega,\tau)$ [22, 23].

The linear prediction (LP) method is also one of the most powerful techniques of analyzing speech signals. LP coefficients have filter characteristics (all-pole type) and the 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 [17, 21]. The LP residue can also be used as a short-term signal $s(t,\tau)$ [19]. Waveform processing and AMDF have also been incorporated [18].

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

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

where $\mathcal{F}^{-1}[\cdot]$ is the Fourier inverse transform, $C_A(q,\tau)$ is the amplitude cepstrum, $C_\phi(q,\tau)$ is the phase cepstrum of $C(q,\tau)$, and q denotes the quefrency (time domain). The complex cepstrum of $X(\omega,\tau)$ in Eq. (7)

can also be represented as

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

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

The amplitude cepstrum, $C_A(q,\tau)$, is generally used in the traditional method so that $C_{A,\mathrm{src}}(q,\tau)$ and $C_{A,\mathrm{flt}}(q,\tau)$ 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. $C_{A,\mathrm{flt}}(q,\tau)$ 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, $C_{A,\mathrm{src}}(q,\tau)$ represents the 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 $C_{A,\mathrm{src}}(q,\tau)$ or to detect periodicity or harmonicity from $C_{A,\mathrm{src}}(q,\tau)$ by eliminating $C_{A,\mathrm{flt}}(q,\tau)$ from $C_A(q,\tau)$. The last processing is referred to as "liftering". Typical approaches are Noll's original method [15] and his clipstrum method [16].

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 $Y(\omega,\tau)$:

$$y(t) = x(t) * h(t) = e(t) * v_\tau(t) * h(t) \quad (10)$$

$$Y(\omega,\tau) = S(\omega,\tau)V(\omega,\tau)H(\omega,\tau) \quad (11)$$

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 a long-term Fourier transform (LTFT) so that the length of analysis (at each τ) should be more than 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 a 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 by Hess (1983) [1] and more recent reviews are those by Hess (1992) and Cheveigné and Kawahara (2001) [2, 3]. A few examples of recent approaches are instantaneous-amplitude [30], instantaneous-frequency [28, 31], and fundamental wave-filtering [33]. There are also comparative evaluations in Atake *et al.*'s (2000), Ishimoto *et al.*'s (2001), and Nakatani and Irino (2004) [2, 3, 28, 29, 31].

We evaluated twelve typical methods to investigate how robust estimates of F_0 were in reverberant environments: ACMWL (AutoCorrelation through Multiple Window-Length) [24], AMDF [9], STFT-ACorrLog (AutoCorrelation of Log-amplitude spectrum on STFT) [22, 23, 13], STFT-ACorrLag (Lag-windowing on STFT) [20], STFT-Comb (Comb filtering on STFT) [11, 12], SHS [14], Cepstrum [15], LPC-residue [19], VFWFF (Voice Fundamental Wave Filtering (Feed-forward type)) [33], TEMPO [25], IFHC (Instantaneous Frequency of Harmonic Components) [28], and PHIA (Periodicity/Harmonicity using Instantaneous Amplitude) [29]. 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.

The characteristics, parameters, and detailed conditions on these algorithms that we used in this comparative evaluation are listed in Table 1. TEMPO was used as a complete original version. As the other methods implemented by the researchers were based on their original research, they were the first or original versions. Several parameters were set to obtain appropriate results based on preliminary simulations. For more detailed information on their technical implementations and skill levels, please refer to their original papers.

3.2 Sound dataset and evaluation measures

The sound dataset we used in this evaluation was the speech database on simultaneous recordings of speech and EGG by Atake *et al.* [28]. 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 we used were created by convolving the original signals, $x(t)$s, with the following reverberant impulse responses, $h(t)$s, as a

Table 1 Characteristics of typical methods of estimating F_0 and their parameter settings

Algorithm	domain	Features	Parameter setting
ACMWL [24]	time	$x(t,\tau)$	Window candidates of nine (15 to 60-ms in 5-ms steps)
AMDF [9]	time	$x(t,\tau)$	40-ms window, 2-ms shift
STFT	freq.		40-ms window, 2-ms shift
auto-corr. [22, 23, 13]	freq.	$\log\lvert X(\omega,\tau)\rvert$	1.5-kHz band-limitation, averaged-clipping
Lag windowing [20]	freq.	$\lvert S(\omega,\tau)\rvert$	3-ms lag windowing, 1.5-kHz band-limitation, averaged clipping
Comb filtering [11, 12]	freq.	$\lvert X(\omega,\tau)\rvert$	10th order of the comb filter
SHS [14]	freq.	$\log\lvert X(\omega,\tau)\rvert$	40-ms window length, 5-ms shift, $N = 15$ (harmonic order), BL = 1.25 kHz (band limitation), weigh function: $h_n = 0.84^{n-1}$
Cepstrum [15]	quef.	$C_A(q,\tau)$	40-ms window, 2-ms shift, 3-ms liftering, Noll's method ($h_w = 0.54 + 0.46\cos(2\pi f/600)$)
LPC Residue [19]	time	$s(t,\tau)$	40-ms window, 2-ms shift, LP order of 12, 1/4 down-sampling
F_0 filtering [33]	time	$s(t,\tau)$	80-ms window, 2-ms shift, 1/4 down-sampling, averaged-clipping, 1.5 kHz band-limitation, "Non-selfsupport mode",
TEMPO [25]	freq.	Fixed point analysis on Instant. freq. (IF)	1-ms F_0 shift length, $nvo = 24$, Heuristic factor: "on"
IFHC [28]	freq.	Harmonicity of IFs	$N_m = 3$ (harmonic order), 1/8 down-sampling, adaptive-windowing based on BW equation
PHIA [29]	time/ freq.	Instant. Amp. (IA) & Dempster's law	400-channel CBFB, 64-channel CQFB, 1/4 down-sampling, averaged-clipping

(1) Hanning windowing function was used in STFT-based processing.

(2) Lower limits of estimated F_0 were set at 60 Hz and higher were set at 800 Hz.

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 the "a" is a gain factor as the normalized power of $h(t)$, T_R is reverberation time, and $n(t)$ is white noise. This is the well-known stochastic-approximated impulse response in room acoustics reported by Schroeder [34, 35], in which the response has an envelope of exponential decay and is a white noise carrier. This formulation for the impulse response has been used in the study of speech intelligibility in room acoustics [36] as general artificial reverberation and thus has non-minimum phase components [34, 35]. 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 within the voiced section have been used as measures for some comparative evaluations in noisy environments [2, 28, 29, 31]. These have concentrated on 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 (%) 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 F_0 and estimated F_0, and this integral is done in the voiced section (t). $N_{F_{0,\text{Est}}}(E)$ is the size of the correct region that satisfies

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

within voiced section (t) where E is the error margin (%). $N_{F_{0,\text{Ref}}}(E)$ is the size of region $F_{0,\text{Ref}}(t)$ in the voiced section at the error margin E. 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 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 approximately indicates gross F_0 error. Since fine F_0 error is the normalized room mean square error between $F_{0,\text{Ref}}(t)$ and $F_{0,\text{Est}}(t)$, 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, T_R. 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

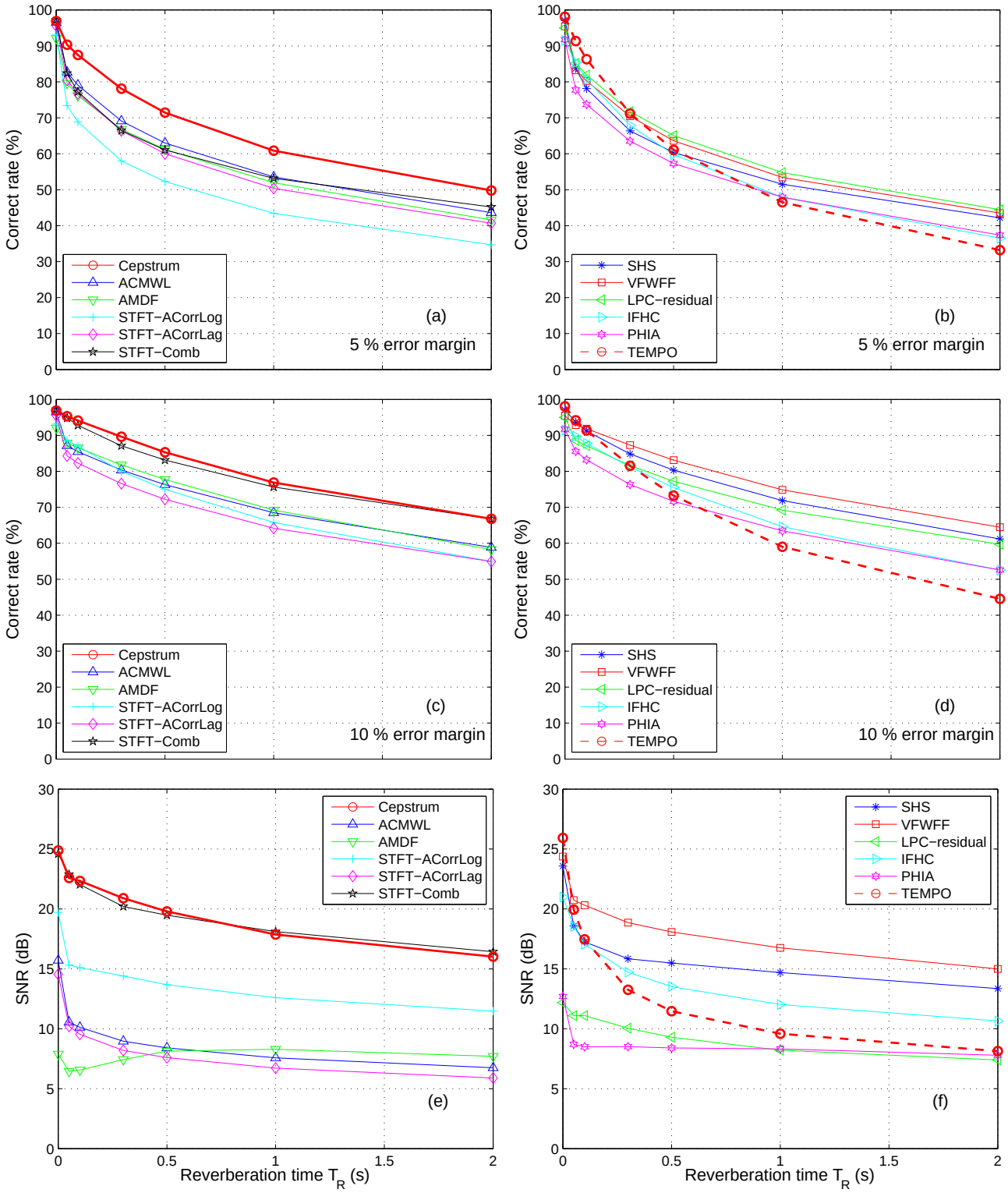

Fig. 2 Estimation results: (a)–(b) percent correct rate within error margin of 5 %, (c)–(d) percent correct rate within error margin of 10 %, and (e)–(f) SNR (s: original, n: error between original and estimated F_0) of F_0 estimates from reverberant speech using twelve typical methods as function of reverberation time, T_R

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 doing 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 more robust against reverberation as we can see from Figs. 2(c) and (e). Maximization of matched harmonicity may have the effect of tracking stationary fluctuations in 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 reverberant speech $y(t)$. From Eqs. (9)-(11), the complex cepstrum of $y(t)$ can be represented as

$$C_Y(q,\tau) = C_{\rm src}(q,\tau) + C_{\rm flt}(q,\tau) + C_H(q,\tau) \quad (16)$$

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").

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

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

where the subscripts "min" and "all" indicate minimum and non-minimum phase characteristics. Here, respective spectra can be represented as

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

where $|X_{\rm all}(\omega,\tau)| = 1$ and $C_{A,{\rm all}}(q,\tau) = 0$. Hence, $C_Y(q,\tau)$ can be separately represented as

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

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

According to Eq. (16), an optimal F_0 estimate is only used to extract $C_{\rm src}(q,\tau)$ from $C_Y(q,\tau)$ to deal with the periodicity/harmonicity of source information as a filter and the reverberation characteristics are eliminated. It is too difficult only to deal with $C_{\rm src}(q,\tau)$ in this estimation task, without measuring $h(t)$ or $C_H(q,\tau)$ (i.e., blind F_0 estimation). In addition, long-term $C_H(q,\tau)$, in which the length of analysis is more than the reverberation time, is needed to accurately extract $C_{\rm src}(q,\tau)$.

We did a preliminary investigation into which component, $C_{H,{\rm min}}(q,\tau)$ or $C_{H,{\rm all}}(q,\tau)$, most affected dealing with $C_{\rm src}(q,\tau)$ for estimating F_0, using Eq. (19). Figure 3 shows the process of 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. Speech signals ($x(t)$ and reverberant $y(t)$) are shown in Figs. 3(a) and (b). The reference F_0 ($F_{0,{\rm Ref}}(t)$ by TEMPO from the EGG signal) and the F_0 ($F_{0,{\rm Est}}(t)$) estimated by the cepstrum method from $y(t)$ correspond to the dashed and solid lines in Fig. 3(c). Note that $F_{0,{\rm Ref}}(t)$ obtained by TEMPO are completely within the voiced part because TEMPO has an Unvoiced/Voiced (U/V) decision. As can be seen, the estimated F_0 was not close to the reference in the voiced part. This method, however, can be used to 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, as plotted in Fig. 3(d). At the same time, two comparative F_0s were obtained as plotted in Figs. 3(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 appears to have a dominant effect from these comparisons of robust and accurate F_0 estimates. Although the same comparisons of all the other stimuli are not presented in this

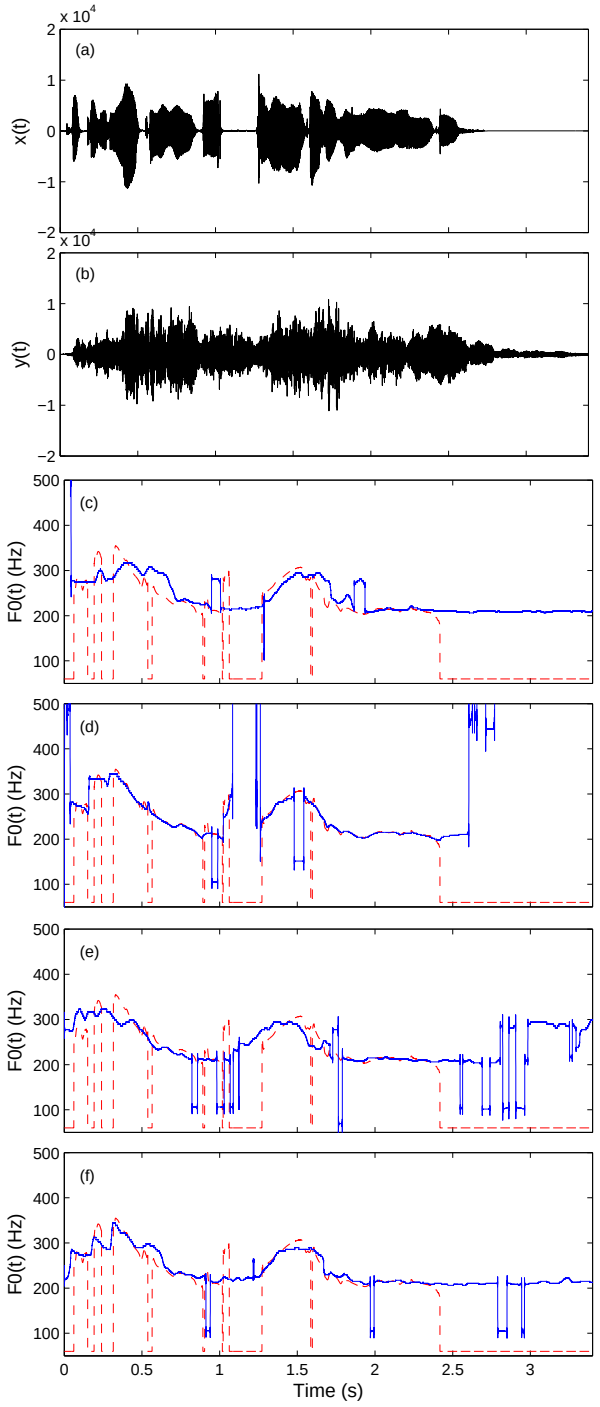

Fig. 3 Examples: (a) original speech $x(t)$, (b) reverberant speech $y(t)$ (reverberation time of 2.0 s), (c) reference F_0 using TEMPO from EGG of $x(t)$ indicated by dashed-line and estimated F_0 using cepstrum method from $y(t)$ indicated by solid line, (d) estimated F_0 from dereverberated $y(t)$ 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

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 with the all-pass component eliminated raised the possibility of achieving robust and accurate estimates of F_0 since we knew homomorphic analysis could easily deal with minimum phase characteristics such as simplified echoes.

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

The MTF concept was proposed by Houtgast and Steeneken [36] 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 [36]. The complex modulation transfer function, $\mathbf{m}(\omega)$, is defined as

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

where $h(t)$ is the impulse response of the room acoustics and ω is the radian frequency. This equation 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

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

where T_R is the reverberation time, i.e., the time required for the power of $h(t)$ to decay by 60 dB [36]. Figure 4 plots the MTF, $m(\omega)$, as a function of the modulation frequency, F_m, (i.e., the dominant frequency in the temporal envelope). These theoretical curves were calculated by substituting five reverberation times (T_R = 0.1, 0.3, 0.5, 1.0, and 2.0 s) and $\omega = 2\pi F_m$ into Eq. (21). Here, $m(\omega)$ can also be regarded as the modulation index with respect to F_m. These curves reveal how much the modulation index of the envelope will be reduced from 1.0 to 0.0 depending on T_R at a specific F_m. In other words, T_R can be predicted from a specific $m(2\pi F_m)$ at a specific F_m. Therefore, the temporal envelope of the reverberant impulse response, $a\exp(-6.9t/T_R)$, can also be predicted using T_R and the "a" in Eq. (13).

Based on the MTF concept, we can establish how much reverberation affects a reduction in $m(\omega)$ and we can then predict the characteristics of room acoustics (T_R) using inverse MTF. MTF-based power-envelope inverse-filtering methods, which have aimed at restoring the reduced MTF in the temporal envelope of the

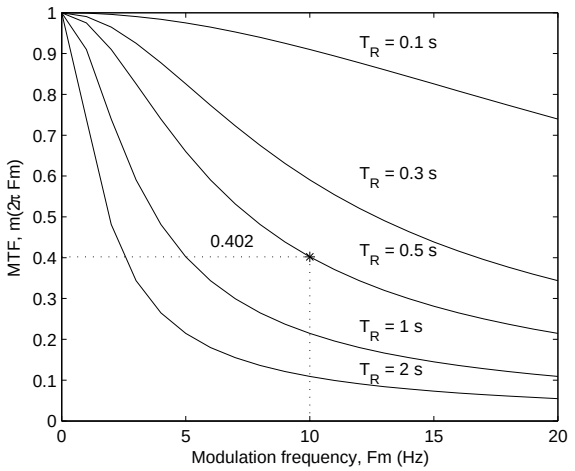

Fig. 4 Theoretical curves representing modulation transfer function $m(2\pi F_m)$ for various conditions with $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s

signals, have been proposed by the present authors [37, 38]. A technique of predicting T_R using the inverse MTF from the temporal envelope of observed signal $y(t)$ has also been proposed [37, 38]; this was used as a constraint for deriving the optimal restoration of the power envelope for the original signal based on the deconvolution relationship between the power envelopes of $y(t)$ and $h(t)$. (For details, see Appendix):

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

where T is the signal duration and $\hat{e}_{x,T_R}(t)^2$ is the set of candidates for the restored power envelope of clean signal $x(t)$ via inverse MTF [37] as a function of T_R. Note that the operation of "$\max(\arg\min\{\cdot\})$" means that 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 [37].

According to Eqs. (12), (13) 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 directly obtained from $\hat{h}(t)$. Although this does not completely equal 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. This is because the MTFs of $h(t)$ and $\hat{h}(t)$ are the same if $\hat{T}_R$ is a complete value, and Eqs. (20) and (21) can be regarded as having characteristics of $C_{H,A}(q,\tau)$ that can be indirectly obtained from $\mathcal{F}^{-1}[\log|\mathbf{m}(\omega)|]$ with the power factor on the LTFT. Therefore, it can be easily predicted that $C_{H,A}(q,\tau)$ becomes a cepstral shape of exponential decay with respect to quefrency (dominant at lower quefrencies).

In contrast, 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 explained 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 basis of LTFT should be done to estimate 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 has not been used in this paper. The 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 out 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 the latter 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 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 that in the higher parts and this exponentially attenuates as the quefrency increases. Therefore, the minimum phase characteristics, $C_{H,\min}(q,\tau)$, can be assumed to concentrate on the lower quefrency parts.

The cepstrum components of the source characteristics are separately concentrated on the higher quefrency parts and those of the filter are separately concentrated on the lower based on the advantages 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).

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

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

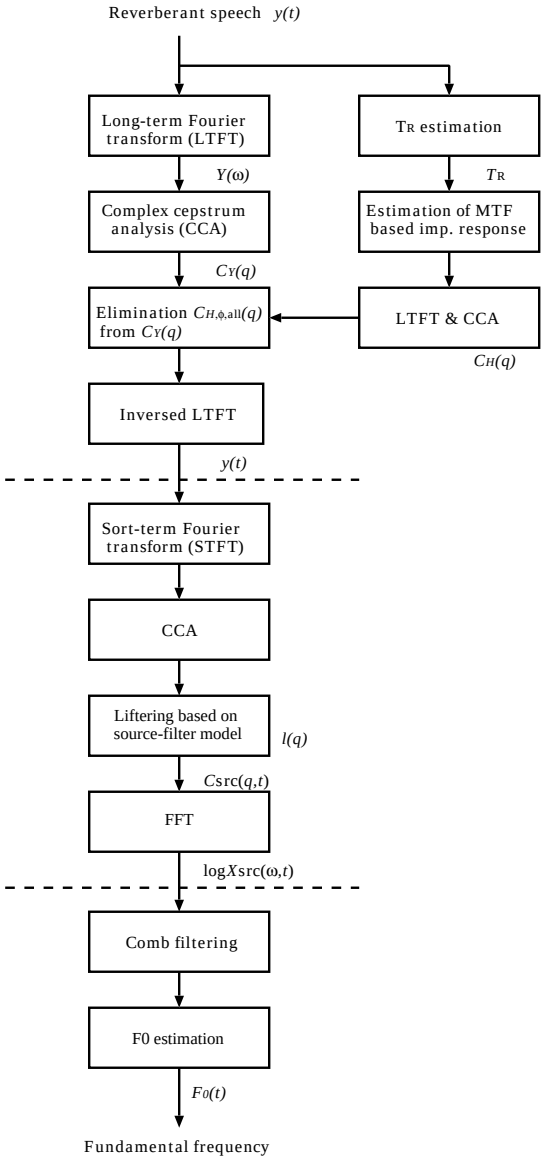

Fig. 5 Algorithm for proposed method

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 by Fig. 5. This method is composed of three main processes: (1) estimating the MTF-based reverberation impulse responses and eliminating them from reverberant speech, (2) extracting $X_{\mathrm{src}}(\omega, \tau)$ 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. 5. As these are commonly used in classical methods of estimation, such as in 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 $C_{\mathrm{src}}(q, \tau)$ can be used to estimate F_0 with the autocorrelation function and/or AMDF. As the aim of this paper was to propose a model concept for robustly estimating F_0 in reverberant environments, these kinds of considerations with regard to modifications in processing are beyond the scope of this paper.

5. Evaluation of 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 out 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 out how effectively $C_{H,\phi,\mathrm{all}}(q, \tau)$ was eliminated on the LTFT with the proposed method.

5.2 Results and discussion

Figure 6 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. 6(a) and (b). Their SNRs are plotted in Fig. 6(c). The results for the cepstrum method indicate the baselines in the evaluations while those for TEMPO (dashed-line) indicate 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 are fewer differences 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 % within an error margin of 5 %, a correct rate of 75 % within an error margin of 10 %, and an SNR of 17 dB at $T_R = 2.0$ s were achieved with 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. 3(b)) are plotted in Figs. 6(d) and (e).

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

concluded that the use of complex cepstrum analysis with regard to non-minimum phase characteristics effectively enabled F_0 to be estimated in reverberant environments.

As shown in Figs. 3(b), 6(d), and 6(e), most F_0-estimation methods including the proposed method, which does not have the U/V decision output of redundant information in the unvoiced and silent sections, while TEMPO does. An important issue is how to determine the U/V decision rules for applications of speech-signal processing in our next stage of research. This issue should be able to be resolved by incorporating power discrimination such as that used in TEMPO with the U/V decision rules, but this is beyond the scope of this paper.

6. Conclusion and Future Perspectives

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 effectively enabled F_0 to be estimated 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. Extensional improvements may be possible by incorporating the proposed method into the U/V decision rules and by considering the robustness of the new method in both noisy and reverberant environments, but these are beyond the scope of this paper.

In future work, we hope to incorporate our new robust and accurate approach to estimate F_0 into the MTF-based process of speech dereverberation [39, 40] and then to improve them so that they can become a more complete blind-dereverberation method. This is because the current method, which consists of MTF-based power-envelope inverse filtering and carrier restoration using F_0 information can dereverberate reverberant speech in power envelopes as well as in resynthesized waveforms, assuming that F_0 can be accurately estimated. As mentioned in the Introduction, robust and accurate F_0 is a significant feature of our new method so that it should be able to contribute to resolving the problems we previously experienced with speech dereverberation [39, 40].

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). It was also partially supported by the Strategic Information and COmmunications R&D Promotion ProgrammE (SCOPE) (071705001) of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

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

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

[3] A. de Cheveigné and H. Kawahara: Comparative evaluation of F0 estimation algorithms, Proc. Eurospeech2001, pp. 2451–2454, Sep. 2001.

[4] 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.

[5] 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.

[6] 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, Apr. 1968.

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

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

[9] 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.

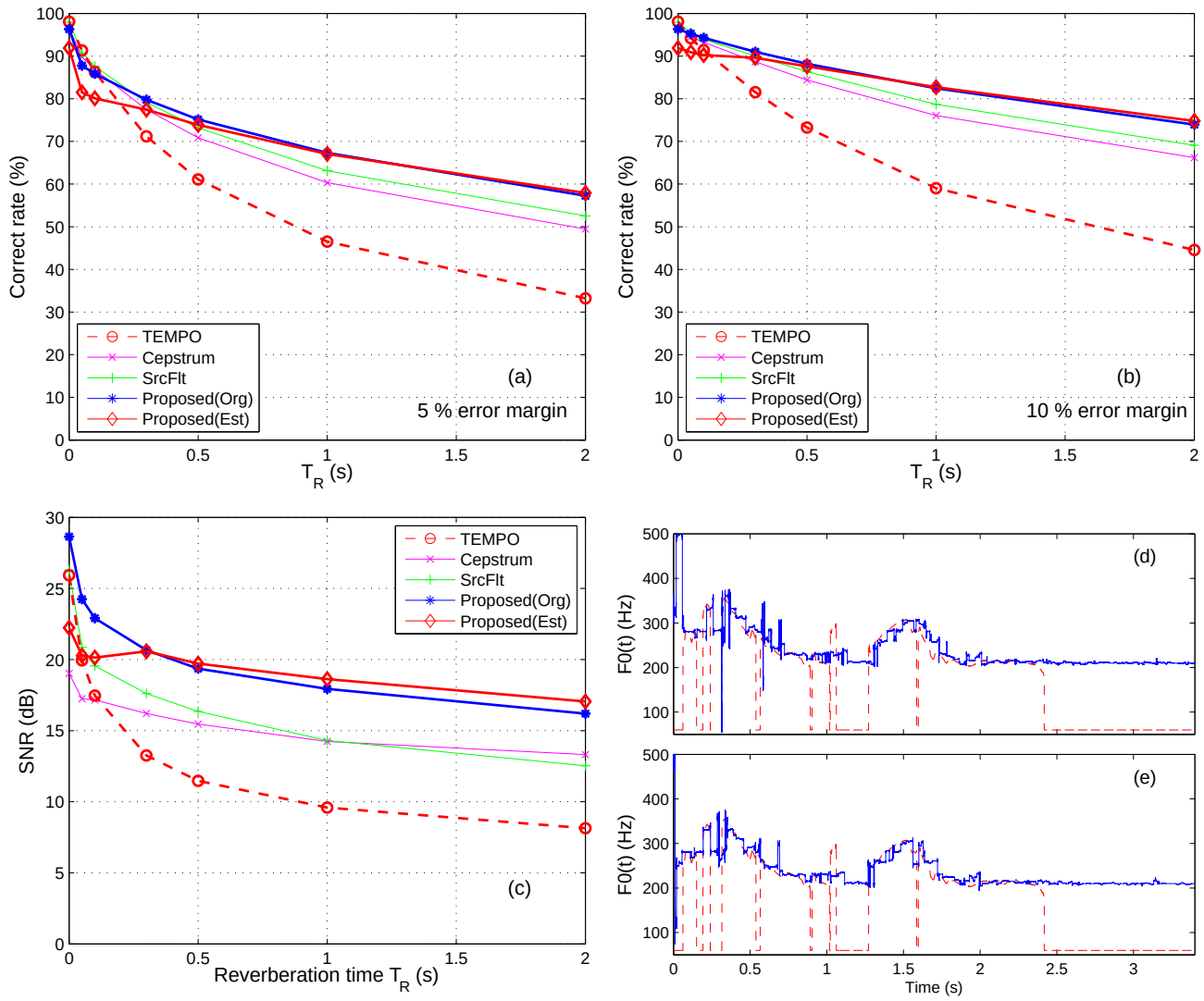

Fig. 6 Evaluation results: (a) percent correct rate within error margin of 5%, (b) percent correct rate within error margin of 10%, (c) SNR of F_0 estimates from reverberant speech using proposed method, and examples of extracted F_0 using proposed model (d) without T_R estimation and (e) with T_R estimation

[10] 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.

[11] 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.

[12] 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, Sep. 1998.

[13] 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.

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

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

[16] A. M. Noll: Clipstrum pitch determination, J. Acoust. Soc. Am., Vol. 44, No. 6, pp. 1585–1591, Jul. 1968.

[17] 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.

[18] 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.

[19] 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.

[20] H. Singer and S. Sagayama: Pitch dependent phone mod-

eling for HMM-based speech recognition, J. Acoust. Soc. Jpn. (E), Vol 15, No. 2, pp. 77–86, Mar. 1994.

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

[22] 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, Mar. 1997.

[23] 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, Jul. 1999.

[24] 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, Sep. 1997.

[25] 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, No. 6, pp. 2781–2784, Sep. 1999.

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

[27] 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, Apr. 1999.

[28] 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, Oct. 2000.

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

[30] 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, Apr. 1999.

[31] 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.

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

[33] 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, Jul. 1995.

[34] M. R. Schroeder: Modulation transfer functions: definition and measurement, Acustica, Vol. 49, pp. 179–182, 1981.

[35] H. Kuttruff: Room Acoustics, Taylor & Francis, fourth edition, London, 2000.

[36] 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).

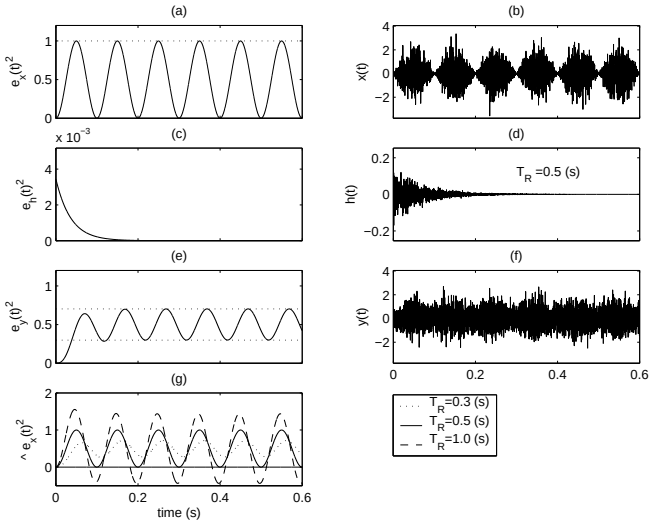

Fig. 7 Examples of relationship between power envelope of system based on MTF concept: (a) power envelope $e_x^2(t)$ of (b) original signal $x(t)$, (c) power envelope $e_h^2(t)$ of (d) impulse response $h(t)$, (e) power envelope $e_y^2(t)$ derived from $e_x^2(t) * e_h^2(t)$, (f) reverberant signal $y(t)$ derived from $x(t)*h(t)$, and (g) restored power envelope $\hat{e}_x^2(t)$

[37] 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, Apr. 2004.

[38] 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, Apr. 2004.

[39] M. Unoki, K. Sakata and M. Akagi: A speech dereverberation method based on the MTF concept, Proc. EuroSpeech2003, pp. 1417–1420, Sep. 2003.

[40] M. Unoki, M. Toi, and M. Akagi: Development of the MTF-based speech dereverberation method using adaptive time-frequency division, Proc. Forum Acusticum 2005, pp. 51–56, Aug. 2005.

Appendix MTF-based power envelope inverse filtering

In the model of inverse-filtering of the power envelope [37], the observed reverberant signal, the original signal, and the stochastic-idealized impulse response in room acoustics are denoted as $bfy(t)$, $\mathbf{x}(t)$, and $\mathbf{h}(t)$, respectively, and are modeled as:

$$\mathbf{y}(t) = \mathbf{x}(t) * \mathbf{h}(t) \quad (24)$$

$$\mathbf{x}(t) = e_x(t)\mathbf{n}_1(t) \quad (25)$$

$$\mathbf{h}(t) = e_h(t)\mathbf{n}_2(t) = a\exp(-6.9t/T_R)\mathbf{n}_2(t) \quad (26)$$

where the asterisks "*" denote the convolution operation, $e_x(t)$ and $e_h(t)$ are the envelopes of $\mathbf{x}(t)$ and

$\mathbf{h}(t)$, and $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ are the mutually independent respective white noise functions, i.e., $< \mathbf{n}_k(t), \mathbf{n}_k(t-\tau) >= \delta(\tau)$. The parameters of the impulse response, a and T_R, correspond to a constant amplitude term and the reverberation time.

In this model, the power envelope of the reverberant signal, $e_y(t)^2$, can be determined as

$$\left\langle \mathbf{y}^2(t) \right\rangle = e_x^2(t) * e_h^2(t) = e_y^2(t) \tag{27}$$

where $< \cdot >$ is the ensemble average operation. This equation shows that there is a significant relationship between the envelopes; i.e., the MTF concept. Based on this result, $e_x^2(t)$ can be recovered by deconvoluting $e_y^2(t)$ with $e_h^2(t)$. Here, the transmission functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are assumed to correspond to the z-transforms of $e_x^2(t)$, $e_h^2(t)$, and $e_y^2(t)$. Thus, the transmission function of the power envelope of the original signal, $E_x(z)$, can be determined from

$$E_x(z) = \frac{E_y(z)}{a^2} \left\{ 1 - \exp\left(-\frac{13.8}{T_R \cdot f_s} \right) z^{-1} \right\} \tag{28}$$

where f_s is the sampling frequency. Finally, the power envelope, $e_x^2(t)$, can be obtained from the inverse z-transform of $E_x(z)$.

Figure 7 shows an example of how power-envelope inverse filtering is related to the MTF concept. Figure 7(a) shows a sinusoidal power envelope as original power envelope $e_x^2(t)$ $(= 0.5(1 + sin(2\pi F_m t))$; the modulation frequency, F_m, was 10 Hz and the modulation index, m, was 1.0). Figure 7(b) shows the original signal, $x(t)$, calculated from $e_x^2(t)$ and a white noise carrier, $n_1(t)$, using Eq. (25). Figure 7(c) shows power envelope $e_h^2(t)$ calculated using Eq. (26) with $T_R = 0.5$ s. Figure 7(d) shows the impulse response, $h(t)$, of Eq. (12), calculated from $e_h^2(t)$ and a white noise carrier, $n_2(t)$. Figures 7(e) and (f) show the power envelope, $e_y^2(t)$, obtained from a convolution of $e_x^2(t) * e_h^2(t)$ and the observed reverberant signal, $y(t)$, obtained from a convolution of $x(t)$ with $h(t)$, respectively. The left panels ((a), (c), and (e)) show the power envelopes of the signals and the right panels ((b), (d), and (f)) show the corresponding signals. In this figure, the modulation index decreased from 1.0 (in Fig. 7(a)) to 0.404 (maximum deviation of envelope between the dotted lines in Fig. 7(e) relative to that in Fig. 7(a)). Since the MTF concept shows the modulation index as a function of F_m and T_R, it can also be shown that the decreased modulation index is derived from $m(2\pi F_m) = 0.402$ using Eq. (21) by substituting $T_R = 0.5$ s and $F_m = 10$ Hz into Eq. (21). The solid line in Fig. 7(g) indicates the restored power envelope $\hat{e}_x^2(t)$ obtained from reverberant power envelope $e_y^2(t)$ (Fig. 7(e)) using Eq. (28) with $T_R = 0.5$ s. We can see that power-envelope inverse filtering can accurately restore the power envelope from a reverberant signal in terms of shape and magnitude.

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.

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).

(Received June 17, 2007; revised September 11, 2007)