EUSIPCO 2013 1569743035

A CEPSTRUM PREFILTERING APPROACH FOR DOA ESTIMATION OF SPEECH SIGNAL IN REVERBERANT ENVIRONMENTS

Ryudo Nagase[†] *Kunio Oishi*[⋆] *Toshihiro Furukawa*[†]

[†] Department of Management Science, Tokyo University of Science
1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
[⋆] School of Computer Science, Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo 192-0982, Japan
Email: nagase-rd@ms.kagu.tus.ac.jp, kohishi@stf.teu.ac.jp, furukawa@ms.kagu.tus.ac.jp

ABSTRACT

A time-quefrency dereverberation method for reducing the minimum-phase component (MPC) of the room impulse response (RIR) is presented. This paper shows that the vocal-tract and glottal components and the MPC of the RIR can be eliminated from observed signals convolving the unknown RIR with an unknown voiced speech signal. The liftered sequence is applied to direction-of-arrival (DOA) estimator with the multiple signal classification (MUSIC) procedure. Computer simulations demonstrate the superiority of the our cepstral prefiltering approach.

Index Terms— Direction-of-arrival (DOA), cepstral prefiltering, time-delay estimation (TDE), cepstral analysis, multiple signal classification (MUSIC) algorithm.

1. INTRODUCTION

The direction-of-arrival (DOA) estimation [1, 2] using microphone array finds application in source localization [3]. The major point of this paper is in DOA estimation from observed signal that is modeled as the convolution of an unknown room impulse response (RIR) with an unknown quasi-stationary source signal. Quasi-stationary signal is modeled as an approximately stationary behavior over short time period. Voiced speech signal is often recognized as quasi-stationary signal. Generally, DOA estimators in reverberant environments are very poor with their estimation errors. Cepstral prefiltering is effective in eliminating the reverberation from received signals [4]. Therefore, it is applied to time-delay estimation (TDE) [4] and the binaural sound localization estimation [3]. In the Stéphenne cepstral prefiltering method, after averaging the minimum-phase component (MPC) of the RIR, a dereverberation is accomplished by subtracting the averaged MPC of the RIR from the observed signal.

This paper presents a dereverberation method using cepstrum analysis for DOA estimation of speech signal in reverberant environments. While the slowly varying vocal-tract and glottal components occur near the origin in the cepstrum, the rapidly varying pitch components occur at a time equal to the pitch period. Moreover, the MPC of the RIR appears near the origin in the cepstrum. In the new approach, short-pass liftering can be used to extract the vocal-tract and glottal components and the MPC of the RIR from the MPC of the observed signal at a time frame. At the same time frame, the pitch period can be liftered from the MPC of the observed signal by a long-pass lifter. To eliminate an additive white Gaussian noise (AWGN) from the short-time liftered component, the average of the short-time liftered component of the observed signal over successive frames is recursively computed. Finally, the all-pass component (APC) of the RIR can be estimated by subtracting the long-pass liftered component and the averaged short-time liftered component from the observed signal. This procedure results in reducing the reverberation from the observed signal. A time delay from the source to a microphone is included in the APC of the RIR. Therefore, if the distance between the source and the microphone is much longer than that between the neighboring microphones, the MUSIC procedure can be applied to DOA estimation. In numerical examples, we show that the new method with the DOA estimator provides a performance better than the conventional method.

2. CONVOLUTIVE MIXING MODEL OF SPEECH AND CEPSTRAL ANALYSIS

We assume that source signal is modeled as quasi-stationary processes. Quasi-stationary process consists of a sequence of variables with mean zero and slowly varying variance over time period. The process presents an approximately stationary behavior over short time interval. Voiced speech signal is often recognized as quasi-stationary signal. The speech is generated by passing either an impulse sequence for voiced speech or a random-noise sequence for unvoiced speech through a slowly time-varying filter of speech production [5]

$$H^{\mathrm{vt}}(z) = \frac{G^{\mathrm{vt}}}{\prod_{i=1}^{L} \left(1 - r_i z^{-1}\right)\left(1 - r_i^{\dagger} z^{-1}\right)}, \qquad (1)$$

where there are $2L$ poles inside the unit circle, that is, $|r_i| < 1$ and the constant G^{vt} is positive. The superscript † denotes conjugation. The speech signal at time n is expressed as

$$s(n) = h^{\text{vt}}(n) * u(n), \tag{2}$$

where $h^{\text{vt}}(n)$ is the inverse z-transform of $H^{\text{vt}}(z)$, the asterisk $*$ denotes time-domain convolution, and $u(n)$ denotes the impulse sequence or the random-noise sequence.

In the convolutive audio mixing model of speech between the source $s(n)$ and J microphones $x_1(n), x_2(n), \cdots, x_J(n)$, assuming that a RIR $h_i^{\text{rir}}(n)$ from the source to the ith microphone without changing over the entire observation interval is a stable and causal non-minimum-phase impulse response, the RIR transformed into the z-transform domain is factored into a MPC and an APC [5, 6]

$$H_i^{\text{rir}}(z) = H_i^{\min}(z) H_i^{\text{ap}}(z). \tag{3}$$

$H_i^{\min}(z)$ represents the MPC

$$H_i^{\min}(z) = G_i \prod_{j=1}^{\infty} \left(1 - a_{ij} z^{-1}\right) \left(1 - a_{ij}^\dagger z^{-1}\right), \tag{4}$$

where $G_i > 0$ and there are zeros inside the unit circle, that is, $|a_{ij}| < 1$. Similarly, $H_i^{\text{ap}}(z)$ represents the APC

$$H_i^{\text{ap}}(z) = \frac{z^{-\tau_i} \prod_{j=1}^{\infty} \left(1 - b_{ij}^{-1} z^{-1}\right) \left(1 - \left(b_{ij}^\dagger\right)^{-1} z^{-1}\right)}{\prod_{j=1}^{\infty} \left(1 - b_{ij} z^{-1}\right) \left(1 - b_{ij}^\dagger z^{-1}\right)}, \tag{5}$$

where there are poles inside the unit circle, that is, $|b_{ij}| < 1$ and zeros outside the unit circle, that is, $|b_{ij}^{-1}| > 1$. If d_i and c represent the distance from the source to the ith microphone and the propagation velocity of the signals respectively, the time delay τ_i is given by

$$\tau_i = \frac{d_i}{c}. \tag{6}$$

We obtain an observed signal at the ith microphone as follows:

$$x_i(n) = h_i^{\text{rir}}(n) * s(n) + n_i(n), \tag{7}$$

where the AWGN $n_i(n)$ with mean zero and variance σ^2 is independent of the source.

The time-domain observed signal is transformed into the frequency domain by the short-time Fourier transform (STFT) as follows:

$$X_i(\omega_k, m) = \sum_{n=0}^{2N-1} w(n) x_i \left\{n + (m-1)T_s\right\} e^{-j\omega_k n}, \tag{8}$$

where T_s is shift size between two neighboring windows, $\omega_k = \pi k/N$ for $k = 0, 1, \cdots, 2N-1$, and $w(n)$ is the exponential window defined as

$$w(n) = \begin{cases} \alpha^n, & 0 \le n < N,\ 0 < \alpha < 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

to reduce the aliasing of the RIR in the complex cepstrum [7]. If $2N$ is significantly larger than the length of the RIR $h_i^{\text{rir}}(n)$, the convolutive mixing model is approximately expressed in the time-frequency (TF) domain as the multiplication

$$X_i(\omega_k, m) \approx H_i^{\text{rir}}(\omega_k) S(\omega_k, m) + N_i(\omega_k, m), \tag{10}$$

where $S(\omega_k, m)$ and $N_i(\omega_k, m)$ are the STFTs of $s(n)$ and $n_i(n)$ at time frame m, and $H_i^{\text{rir}}(\omega_k)$ is the discrete Fourier transform (DFT) of $h_i^{\text{rir}}(n)$. Taking the complex natural logarithm of both side in (10), we obtain

$$\log X_i(\omega_k, m) \approx \log H_i^{\text{rir}}(\omega_k) + \log S(\omega_k, m) + \log V_i(\omega_k, m), \tag{11}$$

where

$$V_i(\omega_k, m) = 1 + \frac{N_i(\omega_k, m)}{H_i^{\text{rir}}(\omega_k) S(\omega_k, m)}. \tag{12}$$

Converting (11) in the quefrency domain by the inverse DFT (IDFT), we have the complex cepstrum

$$\mathcal{X}_i(n, m) \approx \mathcal{H}_i^{\text{rir}}(n) + \mathcal{S}(n, m) + \mathcal{V}_i(n, m), \tag{13}$$

where $\mathcal{H}_i^{\text{rir}}(n)$ is the IDFT of $\log H_i^{\text{rir}}(\omega_k)$. $\mathcal{X}_i(n, m)$, $\mathcal{S}(n, m)$, and $\mathcal{V}_i(n, m)$ are the inverse STFTs (ISTFTs) of $\log X_i(\omega_k, m)$, $\log S(\omega_k, m)$, and $\log V_i(\omega_k, m)$.

$H_i^{\text{rir}}(\omega_k)$ can be factored into the MPC and the APC as follows:

$$H_i^{\text{rir}}(\omega_k) = H_i^{\min}(\omega_k) H_i^{\text{ap}}(\omega_k). \tag{14}$$

The complex cepstrum of (14) is given by

$$\mathcal{H}_i^{\text{rir}}(n) = \mathcal{H}_i^{\min}(n) + \mathcal{H}_i^{\text{ap}}(n), \tag{15}$$

where $\mathcal{H}_i^{\min}(n)$ is zero in the second half of each period:

$$\mathcal{H}_i^{\min}(n) = \begin{cases} 0, & N < n < 2N \\ \mathcal{H}_i^{\text{rir}}(n), & n = 0, N \\ \mathcal{H}_i^{\text{rir}}(n) + \mathcal{H}_i^{\text{rir}}(2N - n), & 0 < n < N \end{cases} \tag{16}$$

and $\mathcal{H}_i^{\text{ap}}(n)$ is an odd function of t:

$$\mathcal{H}_i^{\text{ap}}(n) = \begin{cases} 0, & n = 0, N \\ \mathcal{H}_i^{\text{rir}}(n), & N < n < 2N \\ -\mathcal{H}_i^{\text{rir}}(2N - n), & 0 < n < N. \end{cases} \tag{17}$$

Substituting (15) into (13), we have

$$\mathcal{X}_i(n, m) \approx \mathcal{H}_i^{\min}(n) + \mathcal{H}_i^{\text{ap}}(n) + \mathcal{S}(n, m) + \mathcal{V}_i(n, m). \tag{18}$$

In [4], the MPC of the observed signal $\mathcal{X}_i(n, m)$ in the quefrency domain is given by

$$\mathcal{X}_i^{\min}(n, m) = \begin{cases} 0, & N < n < 2N \\ \mathcal{X}_i(n, m), & n = 0, N \\ \mathcal{X}_i(n, m) + \mathcal{X}_i(2N - n, m), & 0 < n < N. \end{cases} \tag{19}$$

In cepstral processing for speech analysis, calculating the STFT of the windowed speech and taking the complex natural logarithm produces

$$\log S(\omega_k, m) \approx \log H^{\text{vt}}(\omega_k, m) + \log U(\omega_k, m), \quad (20)$$

where $H^{\text{vt}}(\omega_k, m)$ and $U(\omega_k, m)$ are the STFTs of $h^{\text{vt}}(n)$ and $u(n)$ at time frame m. The ISTFT of $\log S(\omega_k, m)$ is the complex cepstrum

$$\mathcal{S}(n, m) = \mathcal{H}^{\text{vt}}(n, m) + \mathcal{U}(n, m). \quad (21)$$

Substituting (21) into (18), we can rewrite (18) as

$$\begin{aligned}\mathcal{X}_i(n, m) \approx & \mathcal{H}_i^{\min}(n) + \mathcal{H}_i^{\text{ap}}(n) + \mathcal{H}^{\text{vt}}(n, m) \\ & + \mathcal{U}(n, m) + \mathcal{V}_i(n, m), \end{aligned} \quad (22)$$

where $\mathcal{H}^{\text{vt}}(n, m)$ and $\mathcal{U}(n, m)$ are the ISTFTs of $\log H^{\text{vt}}(\omega_k, m)$ and $\log U(\omega_k, m)$.

3. A NEW DEREVERBERATION METHOD FOR ESTIMATING DOA

In voiced case, while the vocal-tract and glottal components are slowly varying over successive frames to correspond to the low-time part of the cepstrum due to the exponential window and the all-pole filter, the pitch components are rapidly varying over successive frames to correspond to the high-time part of the cepstrum. Since the values of the vocal-tract and glottal components and the MPC $\mathcal{H}_i^{\min}(n)$ appear near the origin in the cepstrum, we adapt the following short-pass lifter:

$$\begin{aligned}\mathcal{X}_i^{\text{short}}(n,m) &= \begin{cases} \mathcal{X}_i^{\min}(n,m), & 0 \le n < N_{\text{div}} \\ 0, & N_{\text{div}} \le n < 2N \end{cases} \\ &= \begin{cases} \mathcal{H}_i^{\min}(n) + \mathcal{H}^{\text{vt}}(n,m) + \mathcal{V}_i^{\min}(n,m), \\ \qquad\qquad 0 \le n < N_{\text{div}} \\ 0, \quad N_{\text{div}} \le n < 2N, \end{cases} \end{aligned} \quad (23)$$

where $0 < N_{\text{div}} < N_0(m) < N$ and the peak at $N_0(m)$ corresponds to the pitch period in the cepstrum of $\mathcal{X}_i^{\min}(n, m)$. On the other hand, to extract the pitch components, we apply the following long-pass lifter:

$$\begin{aligned}\mathcal{X}_i^{\text{long}}(n,m) &= \begin{cases} 0, & 0 \le n < N_{\text{div}} \\ \mathcal{X}_i^{\min}(n,m), & N_{\text{div}} \le n < 2N \end{cases} \\ &= \begin{cases} 0, \quad 0 \le n < N_{\text{div}} \\ \mathcal{H}_i^{\min}(n) + \mathcal{U}(n,m) + \mathcal{V}_i^{\min}(n,m), \\ \qquad\qquad N_{\text{div}} \le n < 2N. \end{cases} \end{aligned} \quad (24)$$

The average of $\mathcal{X}_i^{\text{short}}(n, m)$ over successive frames to eliminate the AWGN component from $\mathcal{X}_i^{\text{short}}(n, m)$ is recursively computed by

$$\hat{\mathcal{G}}_i^{\text{short}}(n,m) = \begin{cases} \mathcal{X}_i^{\text{short}}(n,m), & m = 1 \\ (1-\mu)\hat{\mathcal{G}}_i^{\text{short}}(n,m-1) + \mu \mathcal{X}_i^{\text{short}}(n,m), & m > 1, \end{cases} \quad (25)$$

where $0 < \mu \le 1$. $\hat{\mathcal{G}}_i^{\text{short}}(n, m)$ is an estimation of $\mathcal{H}_i^{\min}(n) + \mathcal{H}^{\text{vt}}(n, m)$. After reducing the reverberation, the output signal at the ith microphone can be written as

$$\begin{aligned}\mathcal{Y}_i(n, m) \approx & \mathcal{X}_i(n, m) - \hat{\mathcal{G}}_i^{\text{short}}(n, m) - \mathcal{X}_i^{\text{long}}(n, m) \\ \approx & \mathcal{H}_i^{\text{ap}}(n) + \mathcal{Q}_i(n, m), \end{aligned} \quad (26)$$

where

$$\mathcal{Q}_i(n,m) = \begin{cases} 0, & n = 0, N \\ \mathcal{V}_i(n,m), & 0 < n < N_{\text{div}},\ N < n < 2N \\ -\mathcal{V}_i(2N-n,m), & N_{\text{div}} \le n < N. \end{cases} \quad (27)$$

Similarly, in unvoiced case, the all-pole filter is excited by a random-noise sequence. By using the short-pass lifter, we have

$$\begin{aligned}\mathcal{X}_i^{\text{short}}(n,m) &= \begin{cases} \mathcal{X}_i^{\min}(n,m), & 0 \le n < N_{\text{div}} \\ 0, & N_{\text{div}} \le n < 2N \end{cases} \\ &= \begin{cases} \mathcal{H}_i^{\min}(n) + \mathcal{H}^{\text{vt}}(n,m) + \mathcal{U}^{\min}(n,m) + \mathcal{V}_i^{\min}(n,m), \\ \qquad\qquad 0 \le n < N_{\text{div}} \\ 0, \quad N_{\text{div}} \le n < 2N. \end{cases} \end{aligned} \quad (28)$$

The long-pass liftering of $\mathcal{X}_i^{\min}(n, m)$ is given by

$$\begin{aligned}\mathcal{X}_i^{\text{long}}(n,m) &= \begin{cases} 0, & 0 \le n < N_{\text{div}} \\ \mathcal{X}_i^{\min}(n,m), & N_{\text{div}} \le n < 2N \end{cases} \\ &= \begin{cases} 0, \quad 0 \le n < N_{\text{div}} \\ \mathcal{H}_i^{\min}(n) + \mathcal{U}^{\min}(n,m) + \mathcal{V}_i^{\min}(n,m), \\ \qquad\qquad N_{\text{div}} \le n < 2N. \end{cases} \end{aligned} \quad (29)$$

(25) is applied to eliminate the AWGN component from $\mathcal{X}_i^{\text{short}}(n, m)$. We can express the output signal at the ith microphone as

$$\begin{aligned}\mathcal{Y}_i(n, m) \approx & \mathcal{X}_i(n, m) - \hat{\mathcal{G}}_i^{\text{short}}(n, m) - \mathcal{X}_i^{\text{long}}(n, m) \\ \approx & \mathcal{H}_i^{\text{ap}}(n) + \mathcal{P}(n, m) + \mathcal{Q}_i(n, m), \end{aligned} \quad (30)$$

where

$$\mathcal{P}(n,m) = \begin{cases} 0, & n = 0, N \\ \mathcal{U}(n,m), & 0 < n < N_{\text{div}},\ N < n < 2N \\ -\mathcal{U}(2N-n,m), & N_{\text{div}} \le n < N \end{cases} \quad (31)$$

and

$$\mathcal{Q}_i(n,m) = \begin{cases} 0, & n = 0, N \\ \mathcal{V}_i(n,m), & 0 < n < N_{\text{div}}, N < n < 2N \\ -\mathcal{V}_i(2N-n,m), & N_{\text{div}} \le n < N. \end{cases} \quad (32)$$

Thus if $\mathcal{H}_i^{\min}(n)$ and $\mathcal{H}^{\text{vt}}(n, m)$ are rejected from $\mathcal{X}_i(n, m)$ by the new procedure, then $\mathcal{Y}_i(n, m) \approx \mathcal{H}_i^{\text{ap}}(n) + \mathcal{Q}_i(n, m)$ for voiced speech or $\mathcal{Y}_i(n, m) \approx \mathcal{H}_i^{\text{ap}}(n) + \mathcal{P}(n, m) + \mathcal{Q}_i(n, m)$ for unvoiced speech. After taking the STFT, we calculate the complex exponential function for $\log Y_i(\omega_k, m)$.

Table 1. A new cepstrum prefiltering approach for DOA estimation of speech signal.

For all $n = 0, \cdots, 2N-1$ at frame m, do the following:
1) Compute $\mathcal{X}_i^{\min}(n,m)$ as given by (19).
2) Lifter $\mathcal{X}_i^{\min}(n,m)$ by the short-pass lifter.
3) Lifter $\mathcal{X}_i^{\min}(n,m)$ by the long-pass lifter.
4) Compute $\hat{\mathcal{G}}_i^{\text{short}}(n,m)$ recursively using (25).
5) Estimate the APC $\mathcal{H}_i^{\text{ap}}(n)$ as shown in
$\mathcal{Y}_i(n,m) \approx \mathcal{X}_i(n,m) - \hat{\mathcal{G}}_i^{\text{short}}(n,m) - \mathcal{X}_i^{\text{long}}(n,m)$.

$\mathbf{Y}(\omega_k, m)$ forms a column vector by stacking the TF-domain output signal of the J microphones

$$\mathbf{Y}(\omega_k,m) = [Y_1(\omega_k,m), Y_2(\omega_k,m), \cdots, Y_J(\omega_k,m)]^T. \quad (33)$$

Let $\mathbf{R}(\omega_k, m) \in C^{J\times J}$ define the short-time cross-spectral density matrix of the output signal at point (ω_k, m)

$$\mathbf{R}(\omega_k, m) = E\left[\mathbf{Y}(\omega_k, m)\mathbf{Y}(\omega_k, m)^H\right], \quad (34)$$

where $E[\cdot]$ denotes the expectation and the superscript ^{H} denotes conjugate transpose. Calculating the eigenvalues and the corresponding orthonormal eigenvectors of $\mathbf{R}(\omega_k, m)$ produces the noise subspace $\mathbf{U}_n(\omega_k, m)$ orthogonal to the signal subspace $\mathbf{U}_s(\omega_k, m)$ or the column span of $\mathbf{H}^{\text{ap}}(\omega_k)\mathbf{H}^{\text{ap}}(\omega_k)^H$, where

$$\mathbf{H}^{\text{ap}}(\omega_k) = \left[H_1^{\text{ap}}(\omega_k), H_2^{\text{ap}}(\omega_k), \cdots, H_J^{\text{ap}}(\omega_k)\right]^T. \quad (35)$$

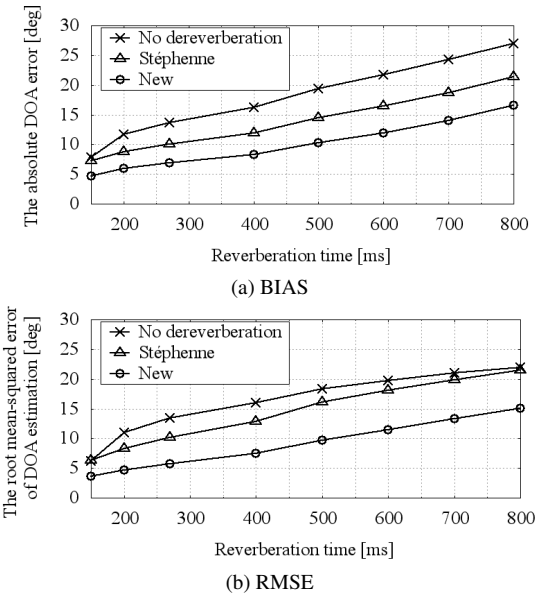
Let $\mathbf{a}(\theta(\omega_k))$ define the steering vector, where d_i is much longer than the distance between neighboring microphones. The source DOA is estimated by finding $\theta(\omega_k)$ such that

$$\mathbf{U}_n(\omega_k, m)^H \mathbf{a}(\theta(\omega_k)) = 0, \quad \theta(\omega_k) \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \quad (36)$$

According to the subspace-based DOA estimation, we can apply the MUSIC procedure [1].

4. SIMULATION RESULTS

The performance of the new method is numerically evaluated by DOA estimation in reverberant environment and we compare it with the conventional Stéphenne dereverberation method. The main objective of this simulation is to evaluate the mean of the absolute DOA error values (BIAS) and the root mean squared errors (RMSE) between the true source directions and their estimated ones obtained from the DOA estimation at each frequency bin except for spatial aliasing with the new and the Stéphenne methods respectively. That is, after reducing the MPC of $H_i^{\text{rir}}(\omega_k)$ from the observed signal by the new or the Stéphenne methods, the MUSIC procedure was applied for DOA estimation. We generated artificial RIRs from a source to two microphones in a room of the size $5.06\times 3.41\times 2.44$ meters using the image method [8] at 96 kHz sampling rate. Thereafter, the artificial RIRs, which are passed through a digital low-pass filter with the stopband edge of 8 kHz are decimated by the factor of 6. The microphones were located at $[2.98, 1.0, 1.5]$ and at $[3.02, 1.0, 1.5]$. The true DOA of the source was $\pm 60°$, $\pm 45°$, $\pm 30°$, and $\pm 15°$ on a two-microphone linear array. The distances from the origin $[3.0, 1.0, 1.5]$ to the sources were 1 meter. Two male speech data and two female speech data were created by catenating independent multiple sentences [9]. The speech data set consisted of 3.26 s long for 1024-point FFT, 3.32 s long for 2048-point FFT, 3.45 s long for 4096-point FFT, 3.71 s long for 8192-point FFT, and 4.22 s long for 16384-point FFT. The exponential window was used for the STFT. The parameter was chosen empirically as the shift size between two neighboring windows of $T_s = 64$, the parameter of averaging $\mathcal{X}_i^{\text{short}}(n,m)$ of $\mu = 0.1$, and the exponential window of $\alpha = 0.994$ for 1024-point FFT, $\alpha = 0.997$ for 2048-point FFT, $\alpha = 0.9985$ for 4096-point FFT, $\alpha = 0.9992$ for 8192-point FFT, and $\alpha = 0.9995$ for 16384-point FFT. In voiced case, since the peak at 5.9 ms corresponded to the averaged pitch period for male, we set N_{div} to 4 ms for male. Meanwhile, since the averaged pitch-period of the female speech was approximately 2.8 ms, N_{div} was set to 2 ms for female. Generally, the averaged pitch period of female speech is shorter than that for male speech. Therefore, N_{div} is set to a value shorter than the averaged pitch-period of female speech, if the source is an unknown speaker being male or female.

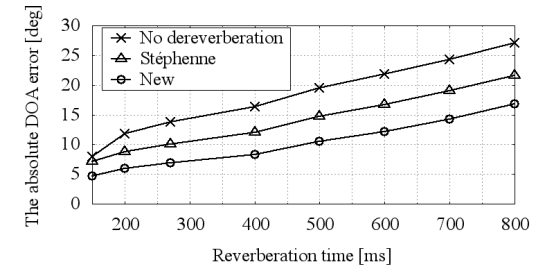


(a) BIAS

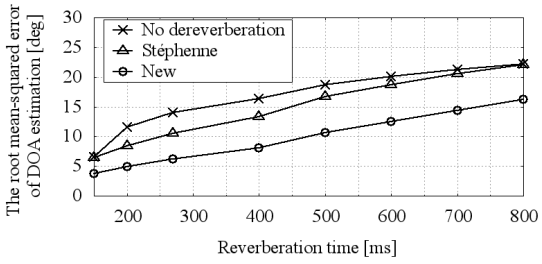
(b) RMSE

Fig. 1. Estimated DOA versus reverberation time for SNR $\approx$ 20 dB, 4096-point FFT, and $N_{\text{div}} = 2$ ms for female speech.

First, we applied the new method in reverberant environments. Comparisons of all procedures with and without the dereverberation method for female speech and for male speech are illustrated in Fig. 1 and in Fig. 2 respectively. These procedures are seen to be essentially biased in reverberant environments. We could achieve a good performance

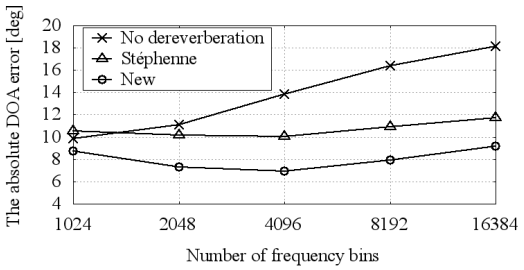


(a) BIAS

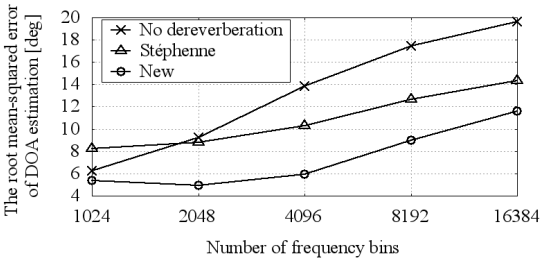


(b) RMSE

Fig. 2. Estimated DOA versus reverberation time for SNR $\approx$ 20 dB, 4096-point FFT, and $N_{\text{div}} = 4$ ms for male speech.



(a) BIAS



(b) RMSE

Fig. 3. Estimated DOA versus number of frequency bins for SNR $\approx$ 20 dB, $N_{\text{div}} = 2$ ms for female, $N_{\text{div}} = 4$ ms for male, and 270-ms reverberation.

for female speech and for male speech. The DOA estimations become worse as reverberation time is longer. The reverberation dominates the DOA estimation error in the long reverberation time range and the slowly time–varying all-pole filter of speech production is negligible. In the long reverberant environments, the DOA error of the proposed method is $4.75°$ and $4.79°$ below the Stéphenne dereverberation method for female speech and for male speech respectively. Therefore, our method is stable in frequency bins as a smoothing mechanism.

Second, it is shown in Fig. 3 how the DOA estimation with the dereverberation methods improves by increasing the number of frequency bins. We expect that in the 270-ms reverberation case, more frequency bins than 2048 are needed to estimate the DOA. As can be seen, a total number of at least 4096 frequency bins are needed to achieve a DOA errors less than 7°. The MUSIC procedure with the new method has the smaller BIAS and RMSE than those with the Stéphenne method.

5. CONCLUSION

We have proposed an approach to remove the MPC of the RIR based on the cepstrum analysis and applied to the DOA estimator. The short-pass lifter operation is a procedure to extract the vocal-tract and glottal components and the MPC of the RIR between the source and the microphone from the observed signal. Moreover, it is useful to extract the pitch components by the long-pass lifter. The experimental results have shown that the MUSIC procedure with the new dereverberation method outperforms the MUSIC with the Stéphenne method in terms of both bias and root mean squared error.

6. REFERENCES

[1] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[2] Wing-Kin Ma, Tsung-Han Hsieh, and Chong-Yung Chi, "DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: a Khatri–Rao subspace approach," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2168–2180, Apr. 2010.

[3] R. Parisi, F. Camoes, M. Scarpiniti, and A. Uncini, "Cepstrum prefiltering for binaural source localization in reverberant environments," *Siginal Processing*, vol. 19, no. 2, pp. 99–102, Feb. 2012.

[4] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Siginal Processing*, vol. 59, no. 3, pp. 253–266, June 1997.

[5] A. V. Oppenheim and R. W. Schafer, "Digital signal processing," *Englewood, Cliffs, NJ: Prentice-Hall*, 1975.

[6] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, Jul. 1979.

[7] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proc. of the IEEE*, vol. 65, no. 10, pp. 1428–1443, Oct. 1977.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[9] "Continuous speech corpus for research," *National Institute of Information*, http://research.nii.ac.jp/src/en/list.html.