MUSIC DEREVERBERATION USING HARMONIC STRUCTURE SOURCE MODEL AND WIENER FILTER

Naoki Yasuraoka,[†‡] *Takuya Yoshioka,*[‡†] *Tomohiro Nakatani,*[‡] *Atsushi Nakamura,*[‡] *Hiroshi G. Okuno*[†]

[†] Graduate School of Informatics
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto 606-8501, Japan

[‡] NTT Communication Science Laboratories
NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun
Kyoto 619-0237, Japan

ABSTRACT

This paper proposes a dereverberation method for musical audio signals. Existing dereverberation methods are designed for speech signals and are not necessarily effective for suppressing long and dense reverberation in musical audio signals because: 1) an all-pole model and a non-parametric model, which are used to represent source spectra, do not match musical tones, and 2) the conventional inverse-filter-based dereverberation is not effective for suppressing long and dense reverberation. To overcome the two problems, an appropriate dereverberation approach for musical audio signals is established. The first problem is resolved by using a harmonic Gaussian mixture model (GMM) to accurately model the harmonic structure of a source spectrum. The second problem is resolved by performing dereverberation with a Wiener filter based on both an estimated inverse filter and an estimated source spectrum model. Experimental results reveal the effectiveness of the proposed dereverberation method using these two solutions.

***Index Terms*—** dereverberation, music signal processing, harmonic GMM, Wiener filter

1. INTRODUCTION

Many musical recordings contain various types of audio effects such as reverberation, delay, and phase shift. Although these audio effects do enhance the perceptual quality of music, they may degrade the automatic analysis of musical audio signals such as melody extraction [1] and chord detection [2]. Therefore, techniques for canceling or controlling such audio effects should be helpful in achieving automatic music analyzers [3] and active music-listening systems [4].

This paper focuses on reverberation among such audio effects. Canceling the effect of reverberation, which is called dereverberation, is an active area of research in speech processing [5, 6], and numerous speech dereverberation methods have been proposed. However, few reports have thus far been published on music dereverberation. We began by applying the existing speech dereverberation method described by Yoshioka et al. [6] to investigate how effective this method was in music dereverberation. A preliminary experiment revealed two problems.

1. Reverberation is not cancelled out very well. The reason for this is twofold. First, the all-pole model used to represent source spectra does not match musical tones. Second, whereas many existing speech dereverberation methods including the considered one are based on blind inverse filtering of room impulse responses, inverse filtering may not achieve accurate dereverberation. This is because the impulse responses of reverberation contained in musical recordings may be extremely long and dense compared to those contained in speech signals and because exact finite-length inverse filters may not exist especially in monaural recordings.
2. The dereverberated signals of struck string instruments sometimes sound like staccato tones. This may be due to difficulty in distinguishing the effect of string vibration from reverberation.

We have considered problem 1 in this paper, and propose a novel dereverberation method suitable for musical audio signals. To address problem 1, two novel ideas are introduced, each of which is aimed at resolving the above-noted two causes of the problem. The first is that a harmonic structure model proposed by Kameoka et al. [7], called the harmonic Gaussian mixture model (harmonic GMM), is used to represent the source spectra. The second is that dereverberation is done by using a Wiener filter that is derived from both of the harmonic structure model and the inverse filter of a room impulse response, which enables to effectively suppress reverberation.

The remainder of this paper is organized as follows: The existing dereverberation method described in [6] is reviewed in Section 2. Section 3 explains the proposed dereverberation method based on harmonic GMM and Wiener filtering. Section 4 presents the evaluation results and Section 5 concludes the paper.

2. REVIEW OF EXISTING DEREVERBERATION METHOD

This section reviews the speech dereverberation method described by Yoshioka et al. [6].

2.1. Problem Statement

First, let us define the task of dereverberation considered in this paper. Let $s(t)$ denote an anechoic signal of speech or music, which we refer to as a source signal. We assume that the source signal is unobservable and that only its reverberated version, denoted by $y(t)$, is available. Dereverberation refers to the process of estimating the source signal, $s(t)$, by using the reverberant signal, $y(t)$. The observed signal is assumed to be monaural.

2.2. Existing Method

The method of [6] works in the STFT domain. Thus, we now denote the STFT coefficients of $s(t)$ and $y(t)$ as $s_{n,l}$ and $y_{n,l}$, where n and l

correspond to the time frame and frequency bin indices, respectively. A statistical model for generating the source signal, $s_{n,l}$, and a reverberation model that transforms $s_{n,l}$ to the reverberated signal, $y_{n,l}$, are introduced. The method first estimates the parameters of these models by maximum likelihood (ML) estimation. Subsequently, the estimated reverberation model is used to estimate $s_{n,l}$ from $y_{n,l}$.

The source and reverberation models are defined as follows:

1. Source signal $s_{n,l}$ independently follows a complex Gaussian distribution in each time frame and frequency bin with mean 0 and variance $\lambda_{n,l}$:

$$s_{n,l} \sim \mathcal{N}_{\mathbb{C}}(s_{n,l}; 0, \lambda_{n,l}). \tag{1}$$

   Note that $\lambda_{n,l}$ corresponds to the short-time power spectral density (psd) of the time-domain source signal at time frame n.

2. The reverberant signal is generated by an auto regressive (AR) system of order K, excited by the source signal as

$$y_{n,l} = \sum_{k=1}^{K} g_{k,l}^{*} y_{n-k,l} + s_{n,l}, \tag{2}$$

   where $g_{k,l}^{*}$ is the k-th AR coefficient at frequency bin l and superscript $*$ stands for complex conjugate. The source signal, $s_{n,l}$, can be recovered by $s_{n,l} = y_{n,l} - \sum_{k=1}^{K} g_{k,l}^{*} y_{n-k,l}$, and hence $[1, -g_{1,l}, \cdots, -g_{k,l}]^{\mathrm{T}}$ may be considered as an inverse filter at frequency l.

In [6], $\lambda_{n,l}$ is modeled by an all-pole model as

$$\lambda_{n,l} = \frac{\nu_n}{\left|1 - \sum_{p=1}^{P} \zeta_{p,n} e^{-j\frac{2\pi l}{L}p}\right|^2}, \tag{3}$$

where $\zeta_{p,n}$ and ν_n correspond to the p-th linear predictor coefficients and prediction residuals at time frame n. Note that $\lambda_{n,l}$ may be represented by using a non-parametric (NP) model as [8]

Based on these assumptions, the probability density function of the whole observed signal, $\mathbf{Y} = \{y_{n,l}\}_{0 \le n \le N-1, 0 \le l \le L-1}$, is expressed as

$$p(\mathbf{Y}; \theta) = \prod_{l=0}^{L-1} \prod_{n=0}^{N-1} \mathcal{N}_{\mathbb{C}}(y_{n,l}; \sum_{k=1}^{K} g_{k,l}^{*} y_{n-k,l}, \lambda_{n,l}), \tag{4}$$

where $\theta = \{\nu_n, \zeta_{p,n}, g_{k,l}\}$ is the set of model parameters. N and L correspond to the number of time frames and frequency bins. Then, the negative log likelihood of θ can be written as

$$\mathcal{L}(\theta) = \sum_{l=0}^{L-1} \sum_{n=0}^{N-1} \left(\log \lambda_{n,l} + \frac{|y_{n,l} - \sum_{k=1}^{K} g_{k,l}^{*} y_{n-k,l}|^2}{\lambda_{n,l}} \right). \tag{5}$$

The θ values are estimated by minimizing $\mathcal{L}(\theta)$ alternately with respect to $\{\nu_n, \zeta_{p,n}\}$ and $\{g_{k,l}\}$. Note that the minimization of $\mathcal{L}(\theta)$ with respect to $\{\nu_n, \zeta_{p,n}\}$ is equivalent to minimizing the Itakura-Saito (IS) divergence between the model psd, $\lambda_{n,l}$, and the source power spectrum estimate given by $|y_{n,l} - \sum_{k=1}^{K} g_{k,l}^{*} y_{n-k,l}|^2$.

After convergence, the dereverberated signal, $\hat{s}_{n,l}$, is calculated as $\hat{s}_{n,l} = y_{n,l} - \sum_{k=1}^{K} \hat{g}_{k,l}^{*} y_{n-k,l}$, where $\hat{g}_{k,l}$ is the estimate of AR coefficient $g_{k,l}$. It should be noted that in [6, 8] the estimated source psd, $\hat{\lambda}_{n,l}$, is not used to calculate $\hat{s}_{n,l}$.

2.3. Limitation of Existing Method

In a severe reverberation situation, Eq. (2) does not accurately represent an actual reverberation process. Hence, the estimated source signal, $\hat{s}_{n,l}$, calculated by inverse filtering as above, contains a significant residual reverberation component due to such a modeling error. Thus, we need to develop a more robust dereverberation scheme against the modeling error.

In addition, the all-pole modeling of source psd $\lambda_{n,l}$ may not be effective for music signals because this model does not represent the harmonic structures appropriately. Thus, we should use another source model suitable for music signals. Although the non-parametric source model [8] can express any power spectra, this model also does not work well in music dereverberation because of its excessive flexibility.

3. DEREVERBERATION METHOD USING HARMONIC GMM SOURCE MODEL AND WIENER FILTERING

3.1. Dereverberation Using Wiener Filter

In order to effectively suppress reverberation, the proposed method uses a Wiener filter instead of the inverse filter to perform dereverberation. The proposed method calculates a dereverberated signal, $\tilde{s}_{n,l}$, as follows:

$$\tilde{s}_{n,l} = W_{n,l} y_{n,l}. \tag{6}$$

$W_{n,l}$ is the Wiener gain defined as

$$W_{n,l} = \frac{\kappa_{n,l}}{\kappa_{n,l} + \gamma |\hat{r}_{n,l}|^2} \quad \text{and} \tag{7}$$

$$\kappa_{n,l} = \alpha \hat{\lambda}_{n,l} + (1 - \alpha) |\hat{s}_{n,l}|^2, \tag{8}$$

where $\hat{r}_{n,l} = \sum_{k=1}^{K} \hat{g}_{k,l}^{*} y_{n-k,l}$, and γ (> 0) and α ($0 \le \alpha \le 1$) are prescribed constants. γ is used to control the amount of reverberation to be suppressed. The motivation for the design of the Wiener filter given by Eqs. (7) and (8) as well as the role of α is as follows. A Wiener filter is generally determined by the power spectra of the source and reverberation signals. These power spectra may be estimated on the basis of the inverse filter coefficient estimate, $\hat{g}_{k,l}$; actually, $|\hat{s}_{n,l}|^2$ and $|\hat{r}_{n,l}|^2$, which are calculated by using $\{\hat{g}_{k,l}\}$, can be used as such estimates. However, $\hat{\lambda}_{n,l}$ also provides a good estimate of the source power spectrum as long as an appropriate source model is used. Whereas $|\hat{s}_{n,l}|^2$ is based on the reverberation model given by Eq. (2), $\hat{\lambda}_{n,l}$ is based on an assumed model for $\lambda_{n,l}$ and is expected to be insensitive to the reverberation modeling error. Therefore, by combining these two source spectrum estimates as in Eq. (8), we can obtain a Wiener filter that is robust against the reverberation modeling error while taking into account the reverberation process given by Eq. (2) at the same time.

3.2. Harmonic GMM Source Model

In order for the above Wiener filter to work effectively, it is essential for $\lambda_{n,l}$ to accurately represent the source spectrum. With this motivation, the harmonic GMM [7] is used to model the source psd $\lambda_{n,l}$ because this model directly expresses the harmonic structures and therefore is more suitable for musical audio signals than the AP model.

The harmonic GMM describes a power spectrum of a musical tone by using a GMM where the means of each Gaussian component appear at the harmonic frequencies of the tone. If we assume J

musical tones, each of which has M harmonics, are present at each time frame, the psd of the m-th harmonics of j-th tone is given by

$$H_{n,l}(j,m) = \frac{v_n(j,m)}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(f(l) - m\mu_n(j))^2}{2\sigma^2}\right], \quad (9)$$

where $v_n(j,m)$ is the relative weight of m-th peak, $\mu_n(j)$ is the fundamental frequency (F0), and σ^2 is the spectral spread of each harmonic component. $f(l)$ means a scaling function that maps the index of frequency bins to Hertz. Therefore, the source psd, $\lambda_{n,l}$, is given by

$$\lambda_{n,l} = \sum_{j=1}^{J}\left(z_n(j)\sum_{m=1}^{M} H_{n,l}(j,m)\right) + \text{Residual}, \quad (10)$$

where $z_n(j)$ corresponds to the intensity of the j-th tone.

The residual component in Eq. (10) represents a noise floor and inharmonic components and is also modeled with a GMM that has I Gaussians with fixed means $\mu^{(I)}(i)$ and fixed large variance $\sigma^{(I)2}$:

$$\text{Residual} = z_n^{(I)}\sum_{i=1}^{I} I_{n,l}(i), \quad (11)$$

$$I_{n,l}(i) = \frac{v_n^{(I)}(i)}{\sqrt{2\pi\sigma^{(I)2}}} \exp\left[-\frac{(f(l) - \mu^{(I)}(i))^2}{2\sigma^{(I)2}}\right]. \quad (12)$$

Here, two restraint conditions are given:

$$\forall n : \sum_{j=1}^{J} z_n(j) + z_n^{(I)} = \sum_{l=0}^{L-1} |s_{n,l}|^2, \quad (13)$$

$$\forall j,n : \sum_{m=1}^{M} v_n(j,m) = 1, \text{ and } \forall n : \sum_{i=1}^{I} v_n^{(I)}(i) = 1. \quad (14)$$

The residual component is not used in the original paper [7], which considers F0 estimation, and is newly introduced in this paper. Without this component, spectral valleys between harmonic peaks become extremely deep, which makes estimation of inverse filter coefficient $g_{k,l}$ unstable. The residual component is aimed at preventing such instability.

3.3. Estimation of Parameters

Estimation of the harmonic GMM parameters is done by minimizing the Kullback-Leibler (KL) divergence between the source psd, $\lambda_{n,l}$, and a power spectrum of the deconvolved signal $\bar{s}_{n,l} = y_{n,l} - \sum_{k=1}^{K} g_{k,l}^* y_{n-k,l}$:

$$\text{minimize} \quad \sum_{n=0}^{N-1}\sum_{l=0}^{L-1} |\bar{s}_{n,l}|^2 \log \frac{|\bar{s}_{n,l}|^2}{\lambda_{n,l}}. \quad (15)$$

Although estimation of source psd parameters is originally formulated as IS-divergence minimization as noted earlier, the harmonic GMM parameters are hard to be analytically optimized according to the IS divergence. Instead, KL-divergence, which often appears in the ML estimation of GMM, yields a simple optimization algorithm. IS-divergence and KL-divergence are not equal but share several important properties such as non-negativity and convexity because: 1) KL-divergence is equivalent to I-divergence where the constraint in Eq. (13) is satisfied, and 2) IS-divergence and I-divergence are in the same group called β-divergence [9]. The definition and relation of these divergences are given in Table 1.

Table 1. Definition of β-divergence.

$\beta \in \mathbb{R}\backslash\{0,1\}$	$\frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right)$	
$\beta = 0$	$\frac{x}{y} - \log\frac{x}{y} - 1$	IS-divergence
$\beta = 1$	$x\log\frac{x}{y} + (y-x)$	I-divergence
$\beta = 2$	$\frac{1}{2}(x-y)^2$	Euclid distance

The source model is then estimated by iteratively updating the model parameters $\{z_n(j), v_n(j,m), \sigma^2, \mu_n(j), z_n^{(I)}, v_n^{(I)}(i)\}$ based on the expectation-maximization (EM) algorithm. The E-step estimates power spectra of individual harmonic components based on the current parameter estimates. For example, the m-th harmonic component of the j-th tone is estimated as

$$\bar{H}_{n,l}(j,m) = \frac{z_n(j)H_{n,l}(j,m)|\bar{s}_{n,l}|^2}{\sum_{j=1}^{J} z_n(j)\sum_{m=1}^{M} H_{n,l}(j,m) + z_n^{(I)}\sum_{i=1}^{I} I_{n,l}(i)}. \quad (16)$$

The M-step updates all the parameters based on the E-step results. For example, the F0 of the j-th tone is updated as

$$\mu_n(j) = \frac{\sum_{m=1}^{M}\sum_{l=0}^{L-1} m f(l)\bar{H}_{n,l}(j,m)}{\sum_{m=1}^{M}\sum_{l=0}^{L-1} m^2\bar{H}_{n,l}(j,m)}. \quad (17)$$

The initial values of $\mu_n(j)$ are estimated using PreFEst(-core) [10] in order to mitigate the problem of local optimum.

4. EXPERIMENTAL EVALUATION

We conducted experiments to evaluate the effectiveness of the proposed method by using both simulated data and real music recordings. The simulation results are reported and discussed in Sections 4.1 and 4.2; the results for real music recordings are presented in Section 4.3.

4.1. Experimental Condition

Nine unaccompanied monophonic musical pieces synthesized with a MIDI tone generator were used as the sources. The musical performances included three violin, three flute, and three cello pieces. Then, the source signals were reverberated by being convolved with two impulse responses, both of which were for musical effect and whose reverberation times RT_{60} were longer than one second. The dereverberation performance was measured by using the log spectral distance improvement (LSDI), which is defined as

$$\text{LSDI} = \text{LSD}(\mathbf{Y},\mathbf{S}) - \text{LSD}(\hat{\mathbf{S}},\mathbf{S}), \quad \text{and} \quad (18)$$

$$\text{LSD}(\boldsymbol{\eta},\boldsymbol{\xi}) \equiv \sqrt{\sum_{n=0}^{N-1}\sum_{l=0}^{L-1}(20\log_{10}\left|\frac{\eta_{n,l}}{\xi_{n,l}}\right|)^2 \Big/ NL}, \quad (19)$$

where $\mathbf{S} = \{s_{n,l}\}_{0\le n\le N-1, 0\le l\le L-1}$ is the true source signal and $\hat{\mathbf{S}} = \{\hat{s}_{n,l}\}_{0\le n\le N-1, 0\le l\le L-1}$ is the dereverberated signal. The other experimental conditions are listed in Table 2.

We used the harmonic GMM, NP, and AP models for the source model, to evaluate the advantages of the harmonic GMM source model over the other two. In this simulation, dereverberation was performed by inverse filtering instead of Wiener filtering. This is because Wiener-filtered results sometimes have deep spectral valleys

Table 2. Experimental conditions.

STFT analysis	Sampling rate	44.1 kHz
	STFT window	1024 pt Gaussian
	STFT shift	256
Parameters	J : Assumed number of musical tones	3
	P : Order of AP model	32

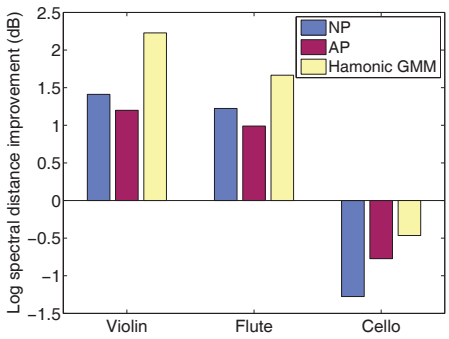

Fig. 1. Experimental results classified in each instrument and each source model.

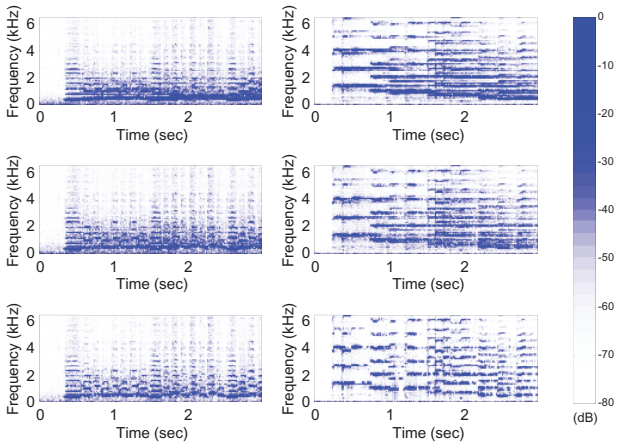

Fig. 2. Spectrograms of reverberant and dereverberated signals of real flute (left) and violin (right) performances. The top panels are of reverberant signals and the middle and bottom panels are of signals dereverberated with the existing and proposed methods.

in a region where the power of the observed signal is low, which degrades the LSDI score, even when the filtered signals are perceptually less reverberant. The effectiveness of Wiener filtering assessed through visual inspection is discussed in Section 4.3.

4.2. Results and Discussion

Fig. 1 shows the LSDI of the three methods for each instrument. The harmonic GMM source model yielded the largest improvement for all instruments, which means the reverberation filter, $g_{k,l}$, can be estimated more accurately by using it. This result indicates the harmonic GMM source model is more effective within the context of musical audio-signal modeling than existing source models.

The cello's LSDIs are negative in all source models. We think there are two reasons for this; the first is that the reverberation model given by Eq. (2) may overfit to suppress the power of the low-frequency band, and the second is that STFT is unsuitable for analysis of low-frequency signals.

4.3. Example for Real Performance Audio Signal

Fig. 2 shows some dereverberation results using audio signals from real performances on commercial CDs. In these examples, α and γ are respectively set to 0.7 and 2, which were the best-performing values for $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\gamma \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$. It can be seen that each musical note appears clearly in the spectrograms obtained with the proposed method, which means that the reverberation component has been effectively removed. In fact, the signals dereverberated by Wiener filtering sounded less reverberant than the signals dereverberated by inverse filtering. All the above results demonstrate the effectiveness of using the harmonic source model and a Wiener filter and the potential of the proposed method.

5. CONCLUSION

This paper has presented a new signal dereverberation method for musical audio signals that are affected by long and dense reverberation. The proposed method uses a harmonic GMM model for accurate estimation of harmonic structures of the source signal, and effectively dereverberates by using a Wiener filter based on both an estimated inverse filter and an estimated source model. Our experimental results revealed that the proposed method could perform music dereverberation more accurately than the existing speech dereverberation method.

Future work will include the improvement of dereverberation for low-frequency signals and struck string instrument signals, and the establishment of criteria for separately evaluating the dereverberation and signal distortion. Extension of our method to multi-channel processing, in particular, stereo input, is also important.

6. REFERENCES

[1] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[2] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 2, pp. 291–301, 2008.

[3] J. A. Moorer, *On the segmentation and analysis of continuous musical sound by digital computer.*, Ph.D. thesis, Stanford University, 1975.

[4] M. Goto, "Active music listening interfaces based on signal processing," in *Proc. ICASSP*, 2007, vol. 4, pp. 1441–1444.

[5] B. Gillespie and L. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," in *Proc. ICASSP*, 2003, vol. 1, pp. 676–679.

[6] T. Yoshioka, T. Nakatani, and M. Miyoshi, "An integrated method for blind separation and dereverberation of convolutive audio mixtures," in *Proc. EUSIPCO*, 2008.

[7] H. Kameoka, T. Nishimoto, and S. Sagayama, "Multi-pitch trajectory estimation of concurrent speech based on harmonic GMM and nonlinear Kalman filtering," in *Proc. Interspeech*, 2004, pp. 2433–2466.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. ICASSP*, 2008, pp. 85–88.

[9] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation (research memo. 802)," Tech. Rep., Institute of Statistical Mathematics, 2001.

[10] M. Goto, "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings," in *Proc. ICASSP*, 2000, vol. 2, pp. 757–760.