Nonparametric Bayesian Dereverberation of Power Spectrograms Based on Infinite-order Autoregressive Processes

Akira Maezawa, Katsutoshi Itoyama, *Member, IEEE*
Kazuyoshi Yoshii, *Member, IEEE* and Hiroshi G. Okuno, *Fellow, IEEE*

***Abstract*—This paper describes a monaural audio dereverberation method that operates in the power spectrogram domain. The method is robust to different kinds of source signals such as speech or music. Moreover, it requires little manual intervention, including the complexity of room acoustics. The method is based on a non-conjugate Bayesian model of the power spectrogram. It extends the idea of multi-channel linear prediction to the power spectrogram domain, and formulates a model of reverberation as a non-negative, infinite-order autoregressive process. To this end, the power spectrogram is interpreted as a histogram count data, which allows a nonparametric Bayesian model to be used as the prior for the autoregressive process, allowing the effective number of active components to grow, without bound, with the complexity of data. In order to determine the marginal posterior distribution, a convergent algorithm, inspired by the variational Bayes method, is formulated. It employs the minorization-maximization technique to arrive at an iterative, convergent algorithm that approximates the marginal posterior distribution. Both objective and subjective evaluations show advantage over other methods based on the power spectrum. We also apply the method to a music information retrieval task and demonstrate its effectiveness.**

***Index Terms*—Dereverberation, Nonparameteric Bayes, Minorization Maximization**

I. Introduction

An audio signal, in the real world, is a mixture of various kinds of source signals, marred by different kinds of reverberation. Reverberation is typically characterized by the *early reflection* component and the *late reverberation* component: the former refers to components of impulse response that contributes to coloration of the spectrum, and the latter refers to components that adds decaying sound. Many studies have tackled the problem of suppressing the reverberation through dereverberation methods, methods for recovering the source signal ("dry" signal) that has driven the reverberant acoustic environment.

We are particularly interested in developing a late reverberation suppression method that works with a wide variety of audio, including speech and music. Such processing is mainly useful for two purposes.

First, it is useful in its own right: the ability to convert a reverberant audio signal to a dry audio signal is a highly useful post-processing technique [1]. Presence of an appropriate amount of reverberation is important for enjoying musical audio, so it is important to be able to adjust the degree of reverberation that is suited to the musical audio. Suppose, for example, that an amateur musician recorded a piece of music in a highly reverberant church, and later found that reverberation smeared out the nuances he/she wanted to convey. With dereverberation techniques, the musician could retrieve the dry signal, change the mixing ratio of the dry signal and reverberation, and emphasize the nuance he wanted to convey. Dereverberation, akin to an equalizer, could also tailor an audio recording to the user's taste. For example, a user might enjoy a more aggressive sound by attenuating the reverberation of a music track, while another user might enjoy a smoother sound by emphasizing the reverberation.

Second, it is a useful front-end to signal recognition task. While recent studies in dereverberation tend to focus on speech recognition tasks [2], we believe that various music information retrieval (MIR) tasks will merit from dereverberation as well. Take, for example, audio-to-score alignment, the task of temporally matching an audio signal to a music score. Audio-to-score alignment is based on matching a spectral time-slice to an audio rendition of the music score (e.g. through the use of a software synthesizer [3], or using a probabilistic model of the audio signal given the music score [4]). Thus, it is vitally important that the rendition of the music score is truly representative of the audio signal that it tries to align. In many musical audio signals, however, the acoustics of a concert hall smears a musician's performance. Therefore, achieving robustness to different acoustics is necessary. Dereverberation may be helpful in this kind of situation, by making the input audio more representative of the music score. We expect other MIR tasks to benefit from dereverberation.

This paper presents a dereverberation method that suppresses the late reverberation of a variety of audio signals,

Manuscript received January 27, 2014; revised August 17; 2014; accepted August 29, 2014. Date of publication **Month Day, Year**. This work was partially supported by Kakenhi 24220006.

A. Maezawa is with Yamaha Corporation, 203 Matsunokijima, Iwata, Shizuoka 430-0942, Japan (e-mail: akira.maezawa@music.yamaha.com). He is also with the Graduate School of Informatics, Kyoto University, Room 412, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan.

K. Itoyama and K. Yoshii are with Graduate School of Informatics, Kyoto University, Room 412, Research Bldg. No.7, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan (e-mail: {itoyama,yoshii}@kuis.kyoto-u.ac.jp).

H. G. Okuno was with Graduate School of Informatics, Kyoto University. He is now with Graduate School of Creative Science and Engineering, Waseda University, 3F, Shinjuku Lambdax Building 2-4-12, Okubo, Shinjuku-ku, Tokyo 169-0072, Japan (e-mail: okuno@aoni.waseda.jp).

EDICS:AUD-SIRR

including music and speech, recorded under a variety of room acoustics. Our goal mandates the following criteria for the design of dereverberation algorithm:

1) Our method should operate using a single-channel input. While there are many stereo recordings, many audio signals are recorded as monaural tracks, e.g. digital re-masters of early LP records or an audio signals recorded on an end-user's smartphone.
2) Our method should be robust to minor variations in the spatial placement of sources that drive the room acoustics, for musical audio has many sources that are close to each other, meaning that each source is convolved by impulse responses with similar magnitude response but different phase response.
3) The model should be robust to various source signals, since musical audio is a mixture of a wide variety of musical instruments, ranging from a whistle to a snare drum.
4) The user should not have to specify the complexity of the reverberation in advance. This means that the method should be robust to various room acoustics, since the room acoustics of many musical audio cannot be known in advance: the acoustics can range from dead-dry (e.g. close-mic), to highly reverberant (e.g. a cathedral).

In order to achieve (1) and (2) simultaneously, we model a reverberant signal on the power spectrogram domain, since a method that operates in the power spectrum domain is relatively insensitive to speaker movements [5].

In order to tackle (3), the constraints on the source signal should be minimal. To this end, we opt to model the sparsity of the source signal, similar to [5], [6]. By sparsity, we mean the mode of the probability density function of each time-frequency bin of the power spectrogram is zero.

Finally, we realize (4) by estimating the reverberation complexity in a data-driven manner. To this end, we seek to use the Dirichlet process (DP) [7], a data-driven approach to infer the model complexity that has found success in various tasks, such as the number of topics contained in a set of documents [8], the number of speakers in a conversation [9], or the number of fundamental frequencies present in a given spectral time-slice [10]. We apply the DP by modeling a reverberant power spectrogram as a mixture of count data, with the mixture component derived from a non-negative auto-regressive (AR) process defined for each frequency bin. Modeling the reverberation as a mixture of count data allows us to use the DP to model the AR coefficients, such that the AR model complexity grows with the complexity of the observed data. By using the DP, the model would introduce additional AR model prediction coefficient when it is sensible to do so. Thus, it infers AR model coefficients that are compact yet expressive enough for the given audio signal. n Section II introduces the readers to existing studies. We introduce our model in Section III, and an inference algorithm in Section IV. We evaluate our method in Section V.

II. Previous Works

There are many ways to partly dereverberate an audio signal, with unique methods tailored for each application. For example, for speech recognition purposes, it is possible to dereverberate the sequence of speech recognition features [11]. For post-production purposes, we seek to recover the dereverberated audio signal. Dereverberation, for this use, may operate either in non-invertible domain, i.e., the transformed representation and the time-domain signal is not bijective, such as the power spectrum, or in invertible domain such as the time domain or the complex spectral domain. Typical dereverberation method focuses on suppressing the late reverberation.

Dereverberation methods on a time-domain signal or complex spectrum seeks to find the transfer function of a room, and recover the dry signal using inverse filtering [12], [13]. When reverberation is assumed to be a moving-average process, and the reverberation filter satisfies a few reasonable assumptions, the problem of source signal estimation becomes that of recovering the noise that drives an auto-regressive system, using a formulation known as the multi-step linear prediction [14]–[16]. By changing the noise model, the reverberation method can be tuned to dereverberate a particular kind of signal. For example, white Gaussian noise [17], auto-regressive model for speech [18]–[22], or mixture of harmonic sounds [23] have been proposed. If one over-specifies the model of the source signal, the method becomes selective. Inversely, if one under-specifies the model, the method becomes more versatile, at the expense of reduced accuracy.

On the other hand, many studies focus on suppressing reverberant signal in the power spectrum domain [5], [24]–[27]. For example reverberation has been formulated as a moving-average process [5], or as a moving-average process with an exponential decay, i.e., as an AR(1) process [6], [27], [28]. In this family of algorithms, the source model can be customized to suit the characteristics of the source signal. To name a few, additive white Gaussian noise [24] or generalized Normal distribution [5] are some of the possibilities.

Dereverberation methods based on the power domain typically assume additivity of the power spectrum, which may degrade the performance. Methods based on the complex spectrum, on the other hand, require no such assumptions, but are highly sensitive to movement of the source signal. To see this, note that the phase response of reverberation changes drastically with a slight displacement of the source signal. However, the magnitude response changes very slightly. Therefore, methods based on the magnitude or power spectrum representation is more robust to source signal displacement than methods based on complex spectrum.

A more general survey of dereverberation that encompasses various topics can be found in [2].

III. Model Formulation

To recap from the introduction, we aim to design a dereverberation method that works with various kinds of audio signals, including music and speech. It models the power spectrogram as a histogram count data, and incorporate prior information that guides the reverberation model towards a good posterior distribution. Namely, it incorporates a sparse prior on the source signal, and uses the DP to model the AR coefficients, allowing the model order of AR to increase without bound as the data mandates.

We develop on the exponential decaying model (AR(1) model) of reverberation in the power spectrogram domain [27], and incorporate ideas from methods formulated in complex spectrum representation [18] to arrive at a higher order AR model – AR(∞), in fact – over the power spectrogram. The infinite limit of the AR order is captured using the DP, and allows the "effective" number of active components in the AR model to be tuned to the input audio signal.

With these in mind, the main contribution of this paper is the extension of the AR(1) model of reverberation in the power spectrum domain [27] to infinite order. First, we justify the higher order AR model, and lay the foundation of the statistical model. Next, we introduce a novel model of reverberant audio that uses the DP to express the reverberation coefficients. Finally, since our model cannot be solved using standard techniques [29], we present a novel posterior inference algorithm based on minorization maximization.

A. Signal Model

In order to model the power spectrogram as a mixture model, we first define what kind of components would comprise the mixture. To this end, we shall extend the multichannel linear prediction (MCLP) [18] to model the observed power spectrogram as a mixture of a source signal and previous observations.

Let $y(f,t) \in \mathbb{C}^{F\times T}$ be the short-time Fourier transform (STFT) of the observation, and $s(f,t) \in \mathbb{C}^{F\times T}$ be that of the source signal. Here, F and T denote the number of frequency bins and the number of frames, respectively. Then, under reasonable assumptions [18], reverberation can be modeled as follows, for some time-invariant filter of order I, $h(f,t) \in \mathbb{C}^{F\times I}$:

$$y(f,t) = s(f,t) + \sum_{i=1}^{I} h(f,i)y(f,t-i). \tag{1}$$

This kind of formulation is known as MCLP, which is originally formulated as a multichannel problem, but is known to work with single-channel audio as well [18]. This formulation ignores early reflection, and takes into account of the late reverberation, whose impulse response is greater than the frame length of the STFT.

We extend MCLP to the power domain. Note that for all (f,i), $h(f,i)$ is zero-mean because its expectation should be invariant to rotations in the complex plane. In other words, where one defines the reverberation to begin should not affect the distribution inherent to the reverberation. Now, let us ignore the short-term correlation typically present in h and assume that for each f, $\{h(f,1),\cdots h(f,I)\}$ are mutually independent. Moreover, assume $s(f,t)$ and $h(f,t)$ are independent for all $t \in [1,T]$. Then, in terms of expectation, the following model of the power spectrogram may be formulated, where $|\cdot|$ denotes the l_2-norm:

$$|y(f,t)|^2 = |s(f,t)|^2 + \sum_{i=1}^{I} |h(f,i)|^2 |y(f,t-i)|^2. \tag{2}$$

Hereon, we shall denote $Y(f,t) = |y(f,t)|^2$, $S(f,t) = |s(f,t)|^2$, and $H(f,i) = |h(f,i)|^2$.

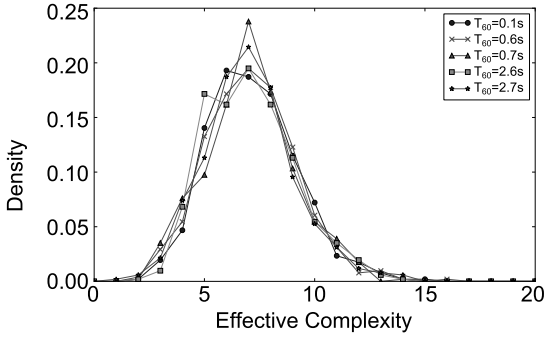


Fig. 1: Distribution of the effective complexity.

Our model can be thought of as a generalization of the exponentially decaying model of reverberation used in [27], where $I = 1$ reduces (2) to an exponential decay model. Note that increasing I makes the reverberation model more expressive but complicates the inference due to the increased number of free parameters. Hence, it is important to choose an I that is compact yet expressive enough for the observed signal.

An investigation suggests that I is best modeled by some finite number that is dependent on the nature of reverberation. Specifically, we determined the AR coefficients of the impulse responses used in Sec. V. A frame length of 1024 samples at 16kHz sampling rate with no overlap was used to compute the power spectrogram of each impulse response. Then, for each impulse response, each frequency bin of the power spectrogram was modeled as a non-negative AR(200) process. The non-negative AR coefficients were estimated using non-negative least squares (NNLS). For each frequency bin, we then sorted the coefficients in descending order, evaluated the cumulative sum, and determined the smallest number for which the cumulative sum exceeds 99% of the total sum. We call this the "effective complexity" at a given frequency bin, in that a significant portion of reverberation is expressed using this number of components.

Fig. 1 shows the density of the effective complexity for different reverberation time (T_{60}, the time it takes for the impulse response of the reverberation to attenuate by 60dB) associated with each impulse response. This figure suggests that the effective complexity is a random variable that is relatively invariant to the kind of reverberation (the reverberation time of the impulse responses ranges from 0 to 2.7 seconds). From this, the figure also suggests that an exponential decay (non-negative AR(1)) model such as [27] may merit from a higher order non-negative AR process.

With this signal model, the fraction of $Y(f,t)$ generated by the source $S(f,t)$, $R_0(f,t)$, is given as follows:

$$R_0(f,t) = \frac{S(f,t)}{S(f,t) + \sum_{i=1}^{I} H(f,i)Y(f,t-i)}. \tag{3}$$

Moreover, the fraction of $Y(f,t)$ generated by the contribution from frame $t-i$, $R_i(f,t)$, is given as follows:

$$R_i(f,t) = \frac{H(f,i)Y(f,t-i)}{S(f,t) + \sum_{i=1}^{I} H(f,i)Y(f,t-i)}. \tag{4}$$

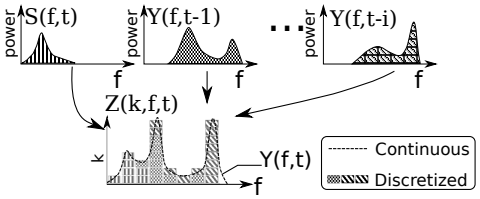


Fig. 2: Conceptual diagram of our statistical model. Power spectrum is treated as a histogram, and each count is associated with a latent component that created it.

These form the basis of the statistical model, which will be discussed next.

B. Statistical Model

We treat a discretized version of $Y(f,t)$ as a histogram count data and formulate a latent variable model that assigns a latent component to each count of the histogram. Such an interpretation are often used either explicitly [30] or implicitly [31] in source separation literatures. It is useful because treating the spectrum as a histogram allows us to apply techniques used in mixture modeling tasks, including the DP. The distributions introduced hereon are defined in Appendix B.

In order to formulate a latent variable model, let us discretize $Y(f,t)$ and set $Y(f,t) := \mathrm{round}(\nu Y(f,t))$ for some $\nu > 0$. We treat $Y(f,t)$ as the number of times a time-frequency bin (f,t) was observed. According to (2), each of $Y(f,t)$ counts in time-frequency bin (f,t) is generated by either the source $S(f,t)$ with a likelihood of (3) or previous observation $Y(f,t-i)$ with a likelihood of (4). The independence of (3) and (4) on the actual count index shows that the counts are independently and identically distributed given (f,t). This concept is illustrated in Fig. 2.

To record the originating component for the kth count in time-frequency bin (f,t), let us introduce $Z_i(k,f,t)$, a one-of-$I+1$ binary variable[1]. "1-of-$I+1$" means that Z is set such that given (k,f,t), exactly one element of $I+1$ choices is 1 and the rest is 0. $Z_0(k,f,t)=1$ indicates that the kth count in time-frequency bin (f,t) originated from the source $S(f,t)$. For $i \in [1,I]$, $Z_i(k,f,t)=1$ indicates that kth count originated from observation from i frames ago, $Y(f,t-i)$. Note that for a given f and t, $Z(k,f,t)$ is defined up to the number of observation at (f,t), that is, $k \in [1, Y(f,t)]$.

Let us re-write Eq. (3) and (4) by introducing, for each f, $\beta(f) \in [0,1]$ and an I-simplex $w(f)$ as follows:

$$\frac{1-\beta(f)}{\beta(f)} = \sum_{i=1}^{I} H(f,i) \tag{5}$$

$$w_i(f) = \frac{H(f,i)}{\sum_{i'=1}^{I} H(f,i')}. \tag{6}$$

[1]In this paper, an index that is normalized to 1 is denoted using subscripts, e.g. $\sum_i Z_i(k,f,t) = 1$ for all (k,f,t). An exception to this convention is the hyperparameters of the prior distribution, which by convention is denoted using a subscript zero. The exceptions to the hyperparameter notation are the variables Z and ϕ, which are zero-indexed variables but are not hyperparameters.

Together, they give the following:

$$R_0(f,t) = \frac{\beta(f)S(f,t)}{N(f,t)} \tag{7}$$

$$R_i(f,t) = \frac{(1-\beta(f))w_i(f)Y(f,t-i)}{N(f,t)} \tag{8}$$

where

$$N(f,t) = \beta(f)S(f,t) + \sum_{i=1}^{I}(1-\beta(f))w_i(f)Y(f,t-i). \tag{9}$$

With this kind of reparametrization, $\beta(f)$ may be thought of as the source signal-to-reverberation ratio for frequency f: the source $S(f,t)$ dominates the observation when $\beta(f)=1$, and the late reverberation dominates the observation when $\beta(f)=0$. $w_i(f)$ may be thought of as the relative contribution of the observation from i frames before, for frequency f.

Equations (7) and (8) respectively indicate the likelihoods that, for a given k, f and t, $Z_0(k,f,t)=1$ and $Z_i(k,f,t)=1$ for $i \in [1,I]$. With these in mind, the conditional posterior of Z given Y is described as follows:

$$p(Z|Y,S,\beta,w) = \prod_{f=1,t=1,k=1}^{F,T,Y(f,t)} \left[\left(\frac{\beta(f)S(f,t)}{N(f,t)} \right)^{Z_0(k,f,t)} \times \prod_{i=1}^{I} \left(\frac{(1-\beta(f))w_i(f)Y(f,t-i)}{N(f,t)} \right)^{Z_i(k,f,t)} \right]. \tag{10}$$

Next, we incorporate prior information on the parameters. Prior information is essential because the maximum likelihood estimation of (10), i.e., maximizing (10) w.r.t. S, Z, β, and w, is susceptible to degenerate solutions. To see why, observe that the maximum value of (10) is 1, for the terms to be exponentiated is at most 1 and the exponent is a binary variable. The maximum likelihood of 1 is attained under many degenerate conditions, such as (1) setting $\beta(f)=1$, (2) setting $\beta(f)=0$ and $w_i(f)=\delta(i-i')$ for some $(i,i') \in [1,I]\times[1,I]$ where $\delta(i)$ is the Kronecker delta, or (3) setting $\beta(f)>0$ and letting $S(f,t) \to \infty$. Therefore, it is critical to incorporate prior on both the source and the reverberation parameters.

1) Prior of w as a Dirichlet Process: Recall that in Section III-A, we observed that I was a random quantity that is dependent on the frequency bin f. Therefore, we seek to convey the idea that the effective complexity grows without bound with the complexity of the observation. We therefore use the DP, which allows a data-driven approach to determine the effective complexity.

To briefly review, a random variable $\bar{X}$ is said to be drawn from a DP over a set $\bar{S}$ with a base distribution $\bar{H}$ and concentration parameter $\bar{\alpha}$ if, for any measurable partition over $\bar{S}$, $\{\bar{A}_i\}_{i=1}^N$, $(\bar{X}(\bar{A}_1),\cdots,\bar{X}(\bar{A}_N)) \sim \mathrm{Dir}\big(\bar{\alpha}\bar{H}(\bar{A}_1),\cdots,\bar{\alpha}\bar{H}(\bar{A}_N)\big)$. GEM (Griffiths, Engen and McCloskey) distribution [32], is a way to represent the distribution of the frequency of each possible value that the base measure emits, and is given as the following infinite-dimensional simplex $\bar{w}$, given infinitely many $\bar{\theta}$:

$$\bar{w}_i = \bar{\theta}(i)\prod_{j=1}^{i-1}(1-\bar{\theta}(j)), \bar{\theta}(i) \sim \mathrm{Beta}\,(1,\bar{\alpha}). \tag{11}$$

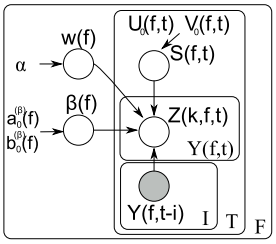


Fig. 3: Graphical model of our method.

$\bar{w}_i$ defines an infinite-dimensional multinomial variable, which can be used as the probability of choosing the ith value. $\bar{w}_i$ can also be thought of as drawing a $\bar{I}$-dimensional Dirichlet distribution $\mathrm{Dir}(\bar{\alpha}/\bar{I})$, as $\bar{I}$ tends to infinity, and permuting it such that the signficant components come first. Because such a Dirichlet distribution is highly sparse, it almost surely outputs a *finite* number of non-negligible components. It tends to emit more non-zero components as $\bar{\alpha}$ increases, or as the observed data becomes so complex that it merits to introduce a new component.

DP has gained popularity in mixture modeling tasks because it specifies the model complexity in a data-driven approach instead of manual specification. Specifically, DP exhibits a "rich-gets-richer" effect [8], where a frequently used component is increasingly likely to be used, and new component becomes less likely to be introduced.

With these in mind, let us apply the DP to model the complexity of reverberation, or the number of effective components in w. Specifically, we let $I \to \infty$, and assume that $w_i(f)$ is an infinite-dimensional simplex drawn from the GEM distribution. That is, for $i \in [1, \infty]$, $w_i(f)$ has a prior of the following form:

$$w_i(f) = \theta(i,f)\prod_{j=1}^{i-1}(1-\theta(j,f)), \theta(i,f) \sim \mathrm{Beta}(1,\alpha)\,. \quad (12)$$

Therefore, we assume that reverberation is a non-negative AR(∞) process, but only a finite number of coefficients contribute to the observation.

2) Prior of S: In order to prevent the degeneracy of (10) by increasing S arbitrarily, we place a sparse prior on S, such that the mode of the distribution is 0. Specifically, we let $S(f,t) \sim \mathrm{Gam}(U_0(f,t), V_0(f,t))$. By setting $U_0 = 1$, the prior of S has a mode at zero, with an exponentially decaying likelihood proportional to $\exp(-S(f,t)/V_0(f,t))$. Moreover, the Gamma distribution is also the simplest distribution that allows for an analytically tractable inference scheme – namely, the sufficient statistics include $-S$ and $\log S$.

The Gamma distribution is chosen because not only is it mathematically convenient but it is also an acceptable posterior distribution for both musical audio and speech signals. We prepared power spectrograms of musical audio signals [33] and speech signals [34], using conditions used in Table I. Then, for each frequency bin, parameters to the Gamma distribution was estimated using maximum-likelihood estimation using 90% of the frames. Finally, Kolmogorov-Smirnov test was performed on the remaining 10% of the dataset to evaluate the goodness-of-fit. We found that the null hypothesis that the underlying distribution of the power spectrogram is distributed as a Gamma was accepted 44% of the time for speech signals, and 67% of the time for musical audio. Therefore, we believe that the Gamma distribution is an acceptable model of the posterior distribution.

3) Prior of β: We let $\beta(f) \sim \mathrm{Beta}\Big(a_0^{(\beta)}(f), b_0^{(\beta)}(f)\Big)$. By setting $a_0^{(\beta)}(f) = b_0^{(\beta)}(f) = 1$ for all f, β becomes non-informative, in that no assumption is made on the signal-to-reverberation ratio, and it is inferred solely from the data. Even though β may cause degenerate solutions as argued previously, we anticipate the sparse prior of S and the prior of w to "compete" with each other to yield in a meaningful posterior distribution. With these in mind, the joint posterior pdf is given as follows, up to a constant normalization factor:

$$\begin{aligned}
&\log p(Z,S,\beta,w|Y,\alpha,\theta,a_0^{(\beta)},b_0^{(\beta)},U_0,V_0) \\
&\quad= \sum_{f=1,t=1,k=1}^{F,T,Y(f,t)} Z_0(k,f,t)\log\left(\beta(f)S(f,t)\right) \\
&\quad+ \sum_{i=1}^{\infty} Z_i(k,f,t)\log\left((1-\beta(f))w_i(f)Y(f,t-i)\right) \\
&\quad- Y(f,t)\log N(f,t) + \log p(S(f,t)|U_0(f,t),V_0(f,t)) \\
&\quad+ \log p(w(f)|\alpha) + \log p(\beta(f)|a_0^{(\beta)}(f),b_0^{(\beta)}(f)). \quad (13)
\end{aligned}$$

The graphical model is shown in Fig. 3. Once the posterior distribution is found, the dry signal may be recovered by multiplying the following time-frequency mask $M(f,t)$ to the power spectrogram:

$$M(f,t) \;=\; \frac{\sum_{k=1}^{Y(f,t)}\langle Z_0(k,f,t)\rangle}{Y(f,t)} \;=\; \left\langle Z_0(\hat{k},f,t)\right\rangle \quad (14)$$

for any $\hat{k} \in [1, Y(f,t)]$. Here, $\langle f(x,y)\rangle_{p(y)}$ denotes the expectation of $f(x,y)$ with respect to a distribution $p(y)$, where $p(y)$ is assumed to be the posterior distribution unless otherwise mentioned.

IV. Marginal Posterior Inference based on MM

In order to compute (14), we seek to marginalize the posterior over all variables but Z. In other words, we want to find the expectation of Z, weighed by all possible combinations of reverberation parameters and source signals. Unfortunately, direct integration $p(Z|Y) = \iiint p(Z,S,\beta,w|Y)dSd\beta dw$ is analytically intractable. Therefore, we will approximate this integral in a manner similar to the Variational Bayes (VB) method, a method to approximate a posterior distribution. In this study, we tighten a constant ($= 0$), using an approximate posterior distribution $q(Z,S,w,\beta)$, which is assumed to be of a factored form ("mean-field" approximation, MFA), as done in VB. Similar approach based on MFA has been employed for discriminative models [35] in the context of conditional random fields. Our model, however, uses MFA to approximate the posterior distribution to a Bayesian latent variable model. To this end, we approximate the posterior $q(Z,S,w,\beta)$ as $\prod_{f,t} q(Z(f,t))q(S(f,t))q(w(f))q(\beta(f))$. Moreover, we consider only up to a finite dimension $\bar{I}$ of $q(w)$. This is

known as the truncation approximation [8]. Note that in this case, the marginal posterior w.r.t. $Z(f,t)$ simply becomes $q(Z(f,t))$. To find an approximate posterior, we optimize the following lower bound[2], obtained using Jensen's inequality, i.e., $\sum_{i=1}^{I} \phi_i f(x_i) \leq f(\sum_{i=1}^{I'} \phi_i x_i)$ for a convex function f and an I'-simplex:

$$\begin{aligned}0 = \log 1 &= \log \int p(S,Z,\beta,w|Y)\mathrm{d}S\mathrm{d}Z\mathrm{d}\beta\mathrm{d}w \\ &\geq \int q(S,Z,\beta,w) \log \frac{p(S,Z,\beta,w|Y)}{q(S,Z,\beta,w)}\mathrm{d}S\mathrm{d}Z\mathrm{d}\beta\mathrm{d}w \\ &= \langle \log p(Z|Y,S,\beta,w)\rangle_{q(S,Z,\beta,w)} - \langle \log q(Z)\rangle_{q(Z)} \\ &\quad + \langle \log p(S|U_0,V_0)\rangle_{q(S)} - \langle \log q(S)\rangle_{q(S)} \\ &\quad + \left\langle \log p(\beta|a_0^{(\beta)},b_0^{(\beta)})\right\rangle_{q(\beta)} - \langle \log q(\beta)\rangle_{q(\beta)} \\ &\quad + \langle \log p(w|\alpha)\rangle_{q(w)} - \langle \log q(w)\rangle_{q(w)}. \end{aligned} \quad (15)$$

The bound is tightened when q is closest, in Kullback-Leibler divergence sense, to the posterior distribution. Note that when no restriction on the functional form of q is assumed, the bound is tightened when $q(S,Z,\beta,w) = p(S,Z,\beta,w|Y)$.

Optimization of (15) w.r.t. $q(Z)$ yields the following:

$$q(Z) = \prod_{f=1,t=1,k=1,i=0}^{F,T,Y(f,t),\bar{I}} \phi_i(f,t)^{Z_i(k,f,t)} \quad (16)$$

where $\phi_i(f,t) = \bar{\phi}_i(f,t) / \sum_{j=0}^{\bar{I}} \bar{\phi}_j(f,t)$, and

$$\log \bar{\phi}_i(f,t) = \begin{cases} \langle \log \beta(f)\rangle + \langle \log S(f,t)\rangle, & i = 0 \\ \langle \log(1-\beta(f))\rangle + \langle \log w_i(f)\rangle + \log Y(f,t-i), & i > 0 \end{cases}. \quad (17)$$

$\phi(f,t)$, or $\langle Z(k,f,t)\rangle$, is independent of k because given f and t, $Z(k,f,t)$'s are distributed i.i.d[3]. From this follows that, since $\sum_{k=1}^{Y(f,t)} \langle Z_i(k,f,t)\rangle = Y(f,t)\phi_i(f,t)$, (14) can be simplified as $M(f,t) = \phi_0(f,t)$.

Updating with respect other variables is complicated since the term $\langle -\log N(f,t)\rangle_{q(S)q(\beta)q(w)}$ in (13) makes our model non-conjugate, meaning that the expectations in (15) cannot be computed analytically. To attack this problem, we optimize a lower-bound to the objective function (15), using minorization-maximization (MM) technique.

4) Design of the minorization function: In order to find an analytical inference rule for the model, we use the MM algorithm [36], also known as auxiliary function method in the context of signal decomposition [5], [37]. The MM algorithm is a convergent algorithm for maximizing an objective function by designing a lower bound to the objective by introducing additional variables (auxiliary variables), and iteratively tightening the bound with respect to the auxiliary variables and the parameter. The key is to design a lower bound such that it is easy to update w.r.t. both the auxiliary variables and the parameter.

[2]Using Jensen's equality to lower-bound the evidence $p(Y)$ instead of 0 yields in a typical formulation of VB.

[3]This is not to say that, for a given (f,t), Z's are identical for all k, i.e., they are constrained to take on the same value: they are separate random variables that have the same expectations.

Inspecting (10) and (13) reveals that inference is complicated by the term $\langle -\log N(f,t)\rangle$ that appears in the log-joint posterior distribution. Computation of the expectation is complicated because the term is cannot be expressed as a weighted sum of the natural parameters of the posterior. Analytical computation of the expectation is possible if we could express it as a weighted sum of the expectations of the natural parameters, $\langle \log \beta(f)\rangle$, $\langle \log(1-\beta(f))\rangle$, $\langle \log w_i(f)\rangle$, $\langle \log S(f,t)\rangle$, and $\langle -S(f,t)\rangle$. Hence, we shall design a lower bound for $\langle -\log N(f,t)\rangle$ that satisfies these properties.

We will introduce a set of auxiliary variables $\mathcal{A}$ to design a lower bound $J(Z,S,w,\beta,\mathcal{A})$, such that maximizing J w.r.t. Z,S,w,β and $\mathcal{A}$ is possible. Our function J shall satisfy the following criteria:

1) $\langle \log p(Y,Z,S,w,\beta)\rangle_q = \max_{\mathcal{A}} J(Z,S,w,\beta,\mathcal{A})$
2) $\langle \log p(Y,Z,S,w,\beta)\rangle_q \geq J(Z,S,w,\beta,\mathcal{A})$
3) $J(Z,S,w,\beta,\mathcal{A})$ is a weighted sum of the expectations of the natural parameters of the posterior, plus $\mathcal{A}$ and a constant.

To simplify the notation, we shall hereon abbreviate the time-frequency index (f,t) when it is obvious, and notate $Y(f,t-i)$ as $Y^{(i)}$.

First, we lower-bound $\langle -\log N(f,t)\rangle$ by a first-order Taylor expansion about $\mathcal{R}(f,t)$:

$$\langle -\log N(f,t)\rangle \geq -\log \mathcal{R}(f,t) - \frac{\langle N(f,t)\rangle - \mathcal{R}(f,t)}{\mathcal{R}(f,t)}. \quad (18)$$

This bound is expressed in terms of a weighted sum of $\log S(f,t)$ and $-S(f,t)$, so we can immediately optimize the objective w.r.t. $q(S)$ to obtain the following:

$$S(f,t) \sim \mathrm{Gam}(U(f,t), V(f,t)) \quad (19)$$

where

$$U(f,t) = U_0(f,t) + Y(f,t)\phi_0(f,t) \quad (20)$$
$$V(f,t) = V_0(f,t) + Y(f,t)\langle \beta(f)\rangle / \mathcal{R}(f,t). \quad (21)$$

The bound is tightened by setting $\mathcal{R}(f,t) = \langle N(f,t)\rangle$.

In order to update (15) w.r.t. $q(w)$ and $q(\beta)$, we set another lower bound to $-\langle N(f,t)\rangle$ used in the right-hand side of (18). Namely, we first bound $-\langle N(f,t)\rangle$ from below by the following bound, for a set of auxiliary variables $\mathcal{V}(f,t) > 0$ and some constant $\mathcal{C}(f,t) > \mathcal{V}(f,t)$:

$$\begin{aligned} -\langle N(f,t)\rangle &\geq \mathcal{V}(f,t) - \mathcal{C}(f,t) - \mathcal{V}(f,t)\log \mathcal{V}(f,t) \\ &\quad + \mathcal{V}(f,t)\left\langle \log\left(\mathcal{C}(f,t) - \langle N(f,t)\rangle_{q(S)}\right)\right\rangle_{q(w)q(\beta)}. \end{aligned} \quad (22)$$

The variable $\mathcal{C}$ can be chosen arbitrarily, provided that $\mathcal{C}(f,t) > \mathcal{C}_{\min}(f,t) = \sqrt{\langle S(f,t)\rangle_{q(S)}^2 + \sum_{i=1}^{\bar{I}} Y(f,t-i)^2}$. Appendix A presents the detailed proof for this bound and the derivation of $\mathcal{C}_{\min}$.

The lower bound of (22) is still not a linear combination of the natural parameters, as it now contains the logarithm of a linear combination of the exponents of natural parameters. Observe, however, that $\log(\mathcal{C}(f,t) - \langle N(f,t)\rangle_{q(S)})$ is expressed as a logarithm of a weighted combination of the exponent of natural parameters. Therefore, we can extract the summation

outside the logarithm by applying Jensen's inequality and obtain the following bound, where $\mathcal{Q}(f,t)$ is a $(\bar{I}+1)$-simplex:

$$\begin{aligned}
&\Big\langle \log(\mathcal{C}(f,t) - \langle N(f,t)\rangle_{q(S)})\Big\rangle_{q(\beta)q(w)} = \\
&\quad\Big\langle \log\Big[\beta(f)\left(\mathcal{C}(f,t) - \langle S(f,t)\rangle_{q(S)}\right) \\
&\quad + \sum_{i=1}^{\bar{I}} (1-\beta(f)) w_i(f) (\mathcal{C}(f,t) - Y(f,t-i))\Big]\Big\rangle_{q(\beta)q(w)} \\
&\geq \mathcal{Q}_0(f,t)\Big\langle \log\Big(\beta(f)\left(\mathcal{C}(f,t) - \langle S(f,t)\rangle_{q(S)}\right)\Big)\Big\rangle_{q(\beta)q(w)} \\
&+ \sum_{i=1}^{\bar{I}} \mathcal{Q}_i(f,t)\Big\langle \log[(1-\beta(f)) w_i(f)\,(\mathcal{C}(f,t) - Y(f,t-i))]\Big\rangle_{q(\beta)q(w)} \\
&\quad - \sum_{i=0}^{\bar{I}} \mathcal{Q}_i(f,t) \log \mathcal{Q}_i(f,t). \quad (23)
\end{aligned}$$

The right-hand side is a weighted sum of the expectations of the natural parameters of β and w used by the joint posterior. Therefore, this bound can be used to analytically optimize the objective function in (15) with respect to $q(\beta)$ and $q(w)$.

Optimizing the function w.r.t. $q(w)$ leads us to the following update (recall w is expressed in terms of θ, by (12)):

$$\theta(i,f) \sim \mathrm{Beta}\Big(a^{(\theta)}(i,f), b^{(\theta)}(i,f)\Big) \quad (24)$$

where

$$\begin{aligned}
a^{(\theta)}(i,f) =& 1 + \xi(i,f) && (25)\\
b^{(\theta)}(i,f) =& \alpha + \sum_{j=i+1}^{\bar{I}} \xi(j,f) && (26)\\
\xi(i,f) =& \sum_{t=1}^{T} Y(f,t)\left(\phi_i(f,t) + \mathcal{Q}_i(f,t)\frac{\mathcal{V}(f,t)}{\mathcal{R}(f,t)}\right). && (27)
\end{aligned}$$

Likewise, optimizing w.r.t. $q(\beta)$ leads to the following:

$$\beta(f) \sim \mathrm{Beta}\Big(a^{(\beta)}(f), b^{(\beta)}(f)\Big) \quad (28)$$

where $a^{(\beta)}$ and $b^{(\beta)}$ are the following:

$$\begin{aligned}
a^{(\beta)}(f) &= a_0^{(\beta)}(f) + \xi(0,f) && (29)\\
b^{(\beta)}(f) &= b_0^{(\beta)}(f) + \sum_{i=1}^{\bar{I}} \xi(i,f) && (30)
\end{aligned}$$

The bound is tightened by updating $\mathcal{Q}$ as follows, and normalizing it such that $\sum_i \mathcal{Q}_i(f,t) = 1$:

$$\begin{aligned}
\mathcal{Q}_0(f,t) \propto& (\mathcal{C}(f,t) - \langle S(f,t)\rangle) \exp\langle \log\beta(f)\rangle \\
\mathcal{Q}_{i>0}(f,t) \propto& (\mathcal{C}(f,t) - Y(f,t-i)) \\
&\times \exp(\langle \log(1-\beta(f))\rangle + \langle \log w_i(f)\rangle). \quad (31)
\end{aligned}$$

5) Expectation of the sufficient statistics: The sufficient statistics are given as follows, where $\psi(x)$ is the digamma function:

$$\begin{aligned}
\langle \log\beta(f)\rangle =& \psi\Big(a^{(\beta)}(f)\Big) - \psi\Big(a^{(\beta)}(f) + b^{(\beta)}(f)\Big) && (32)\\
\langle \log(1-\beta(f))\rangle =& \psi\Big(b^{(\beta)}(f)\Big) - \psi\Big(a^{(\beta)}(f) + b^{(\beta)}(f)\Big) && (33)\\
\langle \beta(f)\rangle =& \frac{a^{(\beta)}(f)}{a^{(\beta)}(f) + b^{(\beta)}(f)} && (34)\\
\langle S(f,t)\rangle =& U(f,t)V(f,t) && (35)\\
\langle \log S(f,t)\rangle =& \log(U(f,t)) + \psi(V(f,t)) && (36)\\
\langle \log w_i(f)\rangle =& \psi\Big(a_i^{(\theta)}(f)\Big) + \sum_{j=1}^{i-1} \psi\Big(b_j^{(\theta)}(f)\Big) \\
&- \sum_{j=1}^{i} \psi\Big(a_j^{(\theta)}(f) + b_j^{(\theta)}(f)\Big) && (37)\\
\langle w_i(f)\rangle =& \frac{a_i^{(\theta)}(f)\prod_{j=1}^{i-1} b_j^{(\theta)}(f)}{\prod_{j=1}^{i}\left(a_j^{(\theta)}(f) + b_j^{(\theta)}(f)\right)}. && (38)
\end{aligned}$$

TABLE I: List of the parameters used.

Parameter	Description	Value
N/A	Waveform Format	mono, 16bit unsigned
N/A	Sampling Frequency	16kHz
F	Frame Length	1024
N/A	Hop Size	25% overlap
N/A	Window function	Modified Bartlett-Hann Window
$a_0^{(\beta)}$, $b_0^{(\beta)}$	Signal-to-Reverb Ratio	1.0, 1.0
U_0 , V_0	Source Sparseness	1.0, 10^{-3}
α	Reverberation Concentration	50.0
$\bar{I}$	Truncation order	50
N/A	Number of iterations of the MM algorithm	20

V. Evaluation

We conduct three experiments to assess the proposed method. First, we compare the performance of our method against two existing methods in the power-spectrum domain, one which relies on source sparsity [5] and another that relies on an AR(1) model of reverberation [27]. Moreover, we explore the behavior of the algorithm when changing a few key parameters. Second, we evaluate our method through a listening test. Finally, we evaluate the effectiveness of dereverberation using audio-to-score alignment, a method in MIR for temporal matching of a piece of musical audio and a music score.

In the following experiments, unless otherwise stated, we use the parameters shown in Table I. The dereverberated signal is created by first multiplying the STFT of the observation $y(f,t)$ and the square-root of the time-frequency mask $M(f,t)$ defined in (14). Then, the time-domain signal is recovered using inverse STFT. This way, the recovered power spectrogram becomes $M(f,t)Y(f,t)$, as specified in (14).

A. Dereverberation performance

1) Data: For the source signals, we used utterances from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus's core test dataset of 24 speakers [34]. For musical sources, we used the BACH10 dataset [33], a database of Bach

chorales recorded in an anechoic chamber. Each signal has a duration of 20 to 30 seconds.

For the impulse response, we prepared real and simulated responses. We prepared five real-world impulse responses, three of which are from the "Impulse response and speech data with microphone array" data from the RWCP Sound Scene Database in Real Acoustic Environment [38], which consist of impulse responses of reverberations whose reverberation time is up to 780ms (33 frames). They include an impulse response of an echo room (denoted "E2A," reverb length of 120ms), a tatami-floored room ("JR1;" 600ms), and a Conference room ("OFC;" 780ms). Here, the length of reverberation is defined as the time that the power of reverberation decays by -60dB compared to the start. The remaining two are from the Promenadikeskus Concert Hall impulse response [39], which contains long reverberation of up to about 3s (130 frames). One impulse response is recorded with a source on stage and mic on stage ("s1_r1_o;" 2.7s); another is recorded with a source on stage and mic on the audience seat ("s1_p1_o;" 2.6s). Therefore, even though they have similar reverberation times, dereverberating the latter is more difficult: the direct-path component is weaker relative to the reverberation, so the source signal is less clear. In other words, the posterior distribution of the variable β is likely to be different, where s1_r1_o is expected to have a smaller $\langle\beta\rangle$ compared to s1_p1_o.

For the synthesized response, the image method [40] was used to synthesize artificial impulse responses. Eighteen impulse responses were generated, with uniform lengths of 20m, heights of 8m, absorption coefficients of 0.19, and widths of 5m to 95m in 5m increment.

2) Metrics: To evaluate our method, we used two metrics: Signal-to-Reverberation Modulation Energy Ratio (SRMR) [41] and the Itakura-Saito Distance (ISD)[4].

SRMR is a non-intrusive measure that increases for signals with shorter late reverberation. Therefore, a dereverberation method should improve the SRMR of the estimated signal from that of the wet signal. The SRMR, however, should not be the only figure of merit because it ignores signal distortion caused by dereverberation: a method may attenuate the late reverberation drastically but sound extremely poorly.

In order to evaluate how close the dereverberated signal sounds to the true dry signal, we evaluate the ISD, an asymmetric measure of spectral dissimilarity defined as follows:

$$\mathrm{ISD}(X, \hat{X}) = \frac{1}{FT} \sum_{f,t} \left(\frac{X(f,t)}{\hat{X}(f,t)} - \log \frac{X(f,t)}{\hat{X}(f,t)} - 1 \right) \quad (39)$$

where $X(f,t)$ is the power spectrogram of the ground-truth dry signal and $\hat{X}(f,t)$ is that of the dereverberated signal. Since we are interested in evaluating the degree to which spectral *shapes* are distorted in the process of dereverberation, the evaluation metric should be invariant to the output gain. Therefore, the estimated signal $\hat{X}$ is scaled by a constant that minimizes the ISD from X to $\hat{X}$. Moreover, we add a small constant $10^{-4} \times \overline{X}$ to both X and $\hat{X}$; because ISD is sensitive to signal power ratios, adding a small floor prevents minute differences in weak signal components from adversely affecting the ISD. Since ISD is a good measure of speech perception [42], we expect ISD to be indicative of the perceptual similarity between the estimated dry signal and the ground-truth dry signal.

[4]Note that there is yet no single validated measure for dereverberation. Therefore, the numerical results presented here should be considered an indication, and not conclusion, on the effectiveness of our method.

TABLE II: Comparison of different methods, averaged over impulse response.

Method	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)
E2A	3.73	6.24	13.01	3.10	3.09
JR1	3.73	6.55	12.94	3.29	3.28
OFC	4.14	6.66	12.30	3.34	3.33
s1_p1_o	5.35	7.53	10.19	3.98	3.93
s1_r1_o	5.12	7.37	9.66	3.96	3.91
Ave.	4.41	6.87	11.62	3.53	3.51

(a) Itakura-Saito Distance.

Method	Dry	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)
E2A	5.01	7.03	9.94	8.80	8.52	8.61
JR1	5.01	4.47	6.52	7.60	5.72	5.83
OFC	5.01	4.60	6.75	7.57	5.88	6.01
s1_p1_o	5.01	3.49	5.40	6.38	4.72	4.88
s1_r1_o	5.01	3.13	4.76	7.55	4.12	4.23
Ave.	5.01	4.54	6.67	7.58	5.79	5.91

(b) Signal-to-Reverberation Modulation Energy Ratio.

We evaluated the data corresponding to the last 10 seconds, since, empirically, all of the methods require some time (about 2s) to settle-in.

3) Comparison with existing methods: We compare our method with two methods that perform dereverberation in the power spectrum domain. The first method, denoted "KAM09," uses a frequency bin-wise convolutive model of reverberation with a sparse source prior [5]. It is similar to our method in that it relies on the sparseness of the source, but different in that it uses a convolutive reverberation model that only incorporates prior on the source signal. The second method, denoted "LEB01," uses a frequency bin-wise AR(1) model of reverberation [27]. It is similar to our method in that it uses an auto-regressive model of reverberation, but different in that it assumes very little on the source signal, incorporates no prior information on the reverberation coefficients, and is restricted to AR(1).

Moreover, we compare our method with a finite-mixture version of our model, which is realized by replacing $\mathrm{GEM}(\alpha)$ (with truncation approximation $\bar{I}$th component) with a finite $\bar{I}$-dimensional $\mathrm{Dir}(\alpha/50)$. This is an approximation of the DP at $\bar{I} = 50$, and the but the expected number of effective component increases with $\bar{I}$ whereas it is fixed in $\mathrm{GEM}(\alpha)$. The inference algorithm is quite similar, and so is omitted; the difference is in the update of $q(w)$, given by (24), and the computation of the sufficient statistics in (37) and (38).

Table II shows the evaluation measures averaged over the source signals. Table IIa shows that our method is capable of decreasing the ISD compared to the wet signal. This is unlike the existing methods, where the ISD *increases* by attenuating the late reverberation. Comparing the infinite version and the finite version of our method suggests that the difference

TABLE III: Evaluation measures averaged over source signals.

	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)
Voice	4.31	6.51	11.41	3.43	3.41
Music	5.35	10.02	13.47	4.46	4.37

(a) Itakura-Saito Distance.

	Dry	Wet	KAM09	LEB01	Proposed (finite)	Proposed (infinite)
Voice	5.11	4.63	6.97	7.78	5.91	6.04
Music	2.06	3.02	3.73	5.86	3.37	3.39

(b) Signal-to-Reverberation Modulation Energy Ratio.

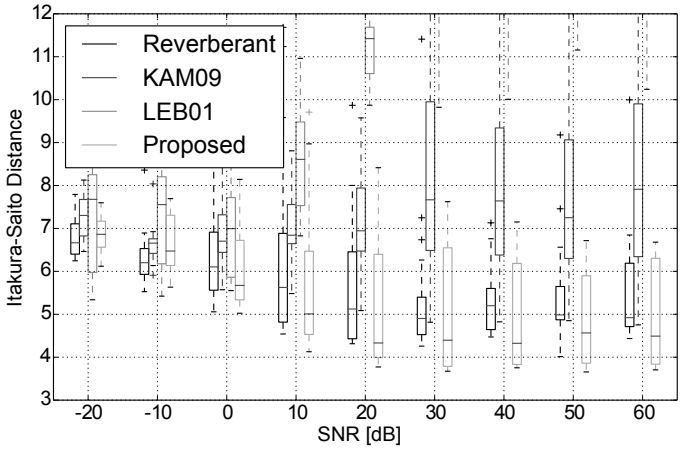
between the two is negligible, in terms of the ISD. Table IIb shows that all the dereverberation methods under consideration are capable of attenuating the late reverberation, since the SRMR increases relative to the wet signal. The data shows that the proposed method is better capable of suppressing the late reverberation compared to the finite version. Even though the existing methods consistently attain higher SRMR than our method, it comes at the price of vastly increased ISD. In this respect, our method is advantageous because its SRMR is comparable to that of the true dry signal, while its ISD is smaller than the wet signal.

Next, we show in Table III the evaluation measures averaged over source signals. Table IIIa shows that our method, again, is capable of making the ISD closer to the true dry signal, and Table IIIb shows that the resulting signal is at least as dry as the true dry signal. The table also shows that the SRMR for music is substantially lower than that of speech, even if they are averaged over the same impulse responses. We believe this owes to the fact that pitched musical instrument sounds are stationary compared to speech, which induces a low-frequency envelope modulation, which in turn decreases the SRMR; in other words, sustained pitched notes sound like reverb.

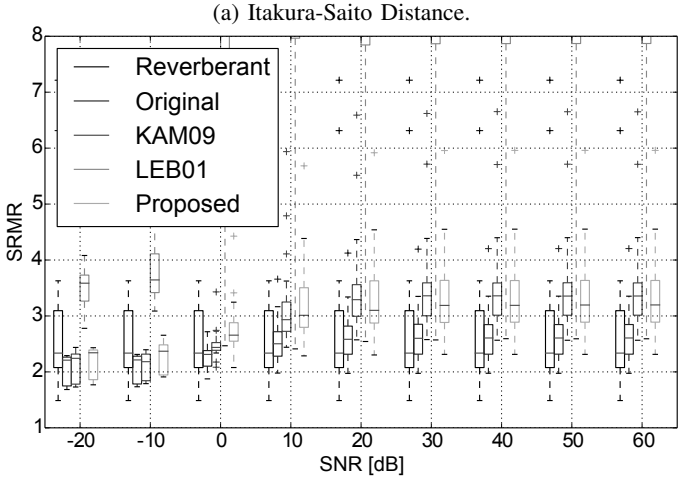
Next, we present the evaluation measures of the synthetic impulse data in Fig. 5. Fig. 5a shows that the ISD increases with the reverberation time, reflecting the intuition that recuperating source signal that has been convolved by a long reverberation is difficult. Fig. 5b shows that the SRMR initially decreases, then increases. In the decreasing region (reverb time$<$1s), early reflection is dominant because the simulated room width is less than the wavelength of the framesize; dereverberation fails in this case because our method only suppresses the late reverberation. In the increasing region, our method starts to suppress late reverberation since late reverberation starts to dominate over early reflection with increased room size.

4) Robustness to noise: We tested the robustness of our method against noise. This is important because our model assumes that the observation is an output from a purely autoregressive system – it ignores the presence of additive noise, such as the LP noise, which is not driven by the AR system that our method tries to estimate.

To this end, the reverberant audio was corrupted by a pink noise, the power spectrum of which rolls off at -10dB per decade. The signal-to-noise ratio (SNR) ranged from -20dB to 60dB, in 10dB increment.



(a) Itakura-Saito Distance.

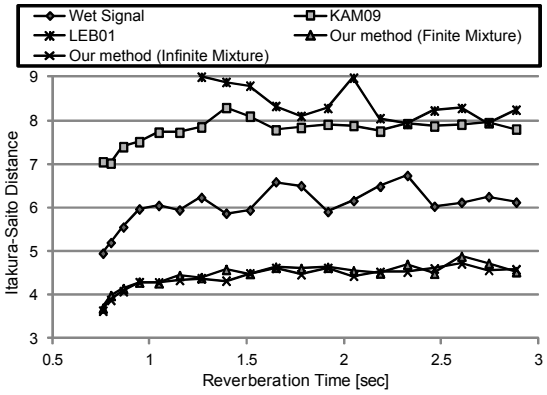


(b) Signal-to-Reverberation Modulation Energy Ratio.

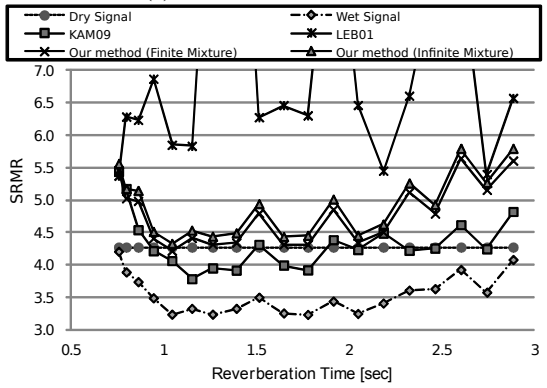
Fig. 4: Evaluation measures with different SNR.

Fig. 4 shows the result. From Fig. 4a, we note that our method fails to dereverberate the signal for a low SNR of less than 0dB. As the SNR increases, the ISD improvement increases, and saturates after about SNR$>$50dB. Our method suffers at a low SNR because the noise masks the wet signal, in particular the weak late reverberation component; the weak late reverberation component provides a valuable cue on higher order AR model parameters, so loss of late reverberation leads to decreased performance. From Fig. 4b, we find that the capability to attenuate the late reverberation saturates after SNR of 10dB or better. These results suggest that our method is admits a weak noise with SNR of about 10dB. This is acceptable for our intended usage of postproduction and preprocessing for MIR tasks: in a typical musical audio, very little noise is introduced after the recording phase.

5) Analysis of key parameters: We evaluate the effect of manually-set room acoustics parameter $\bar{I}$. We have already seen in Table II and III that the infinite model performs better than the finite model, in that the infinite version exhibits roughly the same ISD as the finite version while having higher mean SRMR. This experiment compares the robustness of finite and infinite model by changing $\bar{I}$, which governs the underlying model complexity for the finite case. Recall that the finite version assumes reverberation of AR($\bar{I}$),

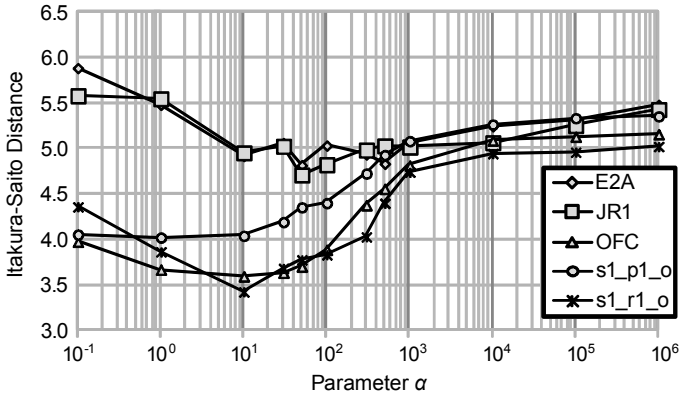


(a) Itakura-Saito Distance.

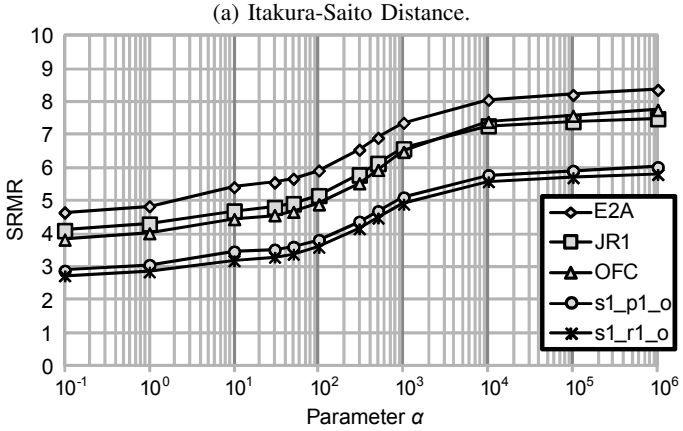


(b) Signal-to-Reverberation Modulation Energy Ratio.

Fig. 5: Evaluation measures of synthetic impulse response as the reverberation time is varied.



(a) Itakura-Saito Distance.



(b) Signal-to-Reverberation Modulation Energy Ratio.

Fig. 6: Evaluation measures as α is varied.

TABLE IV: Standard deviation, evaluated over I, of ISD and SRMR for finite and infinite mixtures, averaged over impulse responses.

	Finite (ISD)	Proposed (ISD)	Finite (SRMR)	Proposed (SRMR)
Voice	0.66	0.42	0.68	0.17
Music	0.12	0.07	0.28	0.14

and the infinite version assumes AR(∞) but evaluates an approximation that computes up to $\bar{I}$. Therefore, we expect the infinite version to be more robust against choice of $\bar{I}$, that is, it is less prone to misspecification of $\bar{I}$. Table IV shows the standard deviation of ISD and SRMR taken for each $\bar{I} \in [32, 48, 64, 96, 128, 192, 256]$. Note that the standard deviation of the infinite model is smaller than the finite model except for SRMR of speech, where the difference is small. This suggests that the infinite model is less sensitive to the manually predetermined complexity of room acoustics, $\bar{I}$, than the finite model. We found that in the finite version, both the SRMR and the ISD tends to increase with increased $\bar{I}$ after a certain point; this suggests that the finite version tends to over-attenuate the reverberation by increasing $\bar{I}$ beyond some optimal point.

Next, we investigate the effect of α, the parameter that governs the tendency for new filter component to be activated. In this experiment, we used only the TIMIT data. To this end, we evaluated the ISD and the SRMR for $\alpha \in [10^{-1}, 10^{6}]$. We show the ISD and SRMR in Fig. 6a and Fig. 6b, respectively. These figures show two tendencies. First, SRMR increases and saturates by increasing α. This matches the intuition that as more components of w are activated, the capability to suppress reverberation tails increase. Second, as α increases, the ISD decreases first and then increases. We believe this behavior is caused by two factors at play. First, as α is increased, the reverberation tail is attenuated, which causes the dereverberated signal to look more like the dry signal. However, after a certain point, the decaying component is attenuated too excessively that the ISD starts to increase. Fortunately, the tendencies of ISD and SRMR as a function of α does not vary too significantly, so very little manual intervention is required for practical purposes.

B. Listening Tests

We performed a subjective evaluation of our method. First, eight 15-second excerpts were prepared. Of eight, three are orchestral excerpts, three are chamber music excerpts (piano solo, piano solo + tenor singer, piano trio), one is a popular music excerpt and one is a male speech. For each excerpt, each subject was first presented, using a headphone, the original reverberant signal. Then, the dereverberated signal using KAM09 and the proposed method were presented in

TABLE V: Average rating (on a 5 point scale) of excerpts averaged over the musical genre. Number following $\pm$ is the standard deviation.

	KAM09	Proposed
Orchestra	1.72±0.74	3.40±0.81
Chamber Music	1.64±0.67	3.74±0.76
Popular Song	1.70±0.74	3.17±0.98
Speech	2.00±1.13	3.17±1.04

random order[5]. Finally, the subjects were asked to rate the sound quality on a 1–5 scale. LEB01 was not used because Table II suggests that KAM09 is preferable over LEB01 both in terms of ISD and SRMR. The test subjects (17 subjects) were all male between the age range of 20 to 40, and deal with sound signal processing technology on a daily basis.

Table V shows the average rating and the standard deviation. The data seems to be in favor of our algorithm. To investigate this, two-way ANOVA was performed on the collected ratings data. With a significance level of 0.01, interaction between the excerpt and deverberation method was not seen (p-value of 0.02), nor did the excerpt itself (p-value of 0.29), but the effect of dereverberation method was significant (p-value of 1.1^{-16}). This suggests that our method produces a more natural sounding dereverberated signal than the existing method[6].

C. Application to MIR Tasks – Audio-to-score Alignment

Audio-to-score alignment is a method to temporally align a music score (e.g. standard MIDI file (SMF)), and an audio recording that plays the music score. It has uses in applications such as score-aided source separation [43], or automated page turning [44].

We use a score alignment method based on [4]. The method is based on fitting a sum of sinusoids to the observed power spectrogram, by interpreting the spectrogram as a histogram. Such interpretation dovetails with our treatment of power spectrogram.

We synthesize three kinds of audio signals. First, ten SMF files from the RWC Classical Music Database [45] are used to synthesize a musical audio signal with a synthesizer, along with a mapping of ground truth score position to audio position. The instrumentations of the chosen SMFs are duets of a single instrument and a piano. Hence, the SMFs are a polyphonic mixture of both sustained and decaying instrumental sound. Next, another set of audio signal is created by convolving the synthesized audio with a reverberation [39] (s1_r1_o). Finally, another set of audio signal is created by dereverberating the synthesized audio with reverb.

We align the music score to three kinds of audio signals, and compare the alignment error percentile. The result is shown in Fig. 7. We first find that audio reverberation multiplicatively corrupts the alignment. At the same time, we also find that dereverberation brings the performance of score alignment to the same level as the ideal case, where no reverberation is added. The dereverberated signal sometimes performs better than the ideal case presumably because our method suppressed not only the reverberation but also sustain pedals of the piano parts; sustain pedal causes the piano string to vibrate freely, elongating duration of the played notes than that notated in the music score.

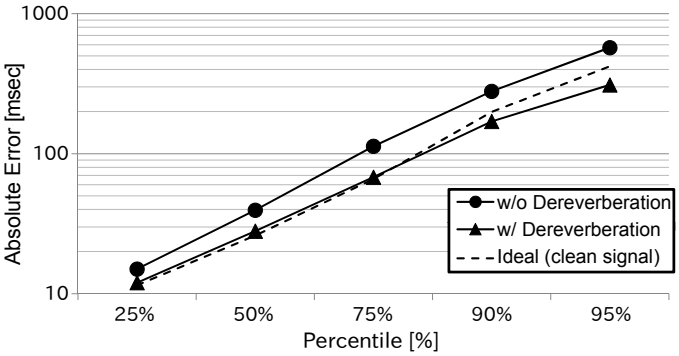


Fig. 7: Percentile of Audio-to-Score Alignment Error.

Dereverberation is clearly beneficial for improving score alignment accuracy. This is particularly beneficial for score-aided source separation, where the fundamental assumption is that the music score and the audio signal is aligned perfectly.

VI. Conclusion

This paper presented a dereverberation method based on the power-spectrogram representation of audio. The method developed on the frequency-dependent auto-regressive model of reverberation to arrive at a statistical model of non-negative auto-regressive model of reverberation based on the DP, such that the effective number of component grows without bound as the data mandates it. Moreover, it incorporated sparse source prior and a non-informative source-to-reverberation ratio prior.

We formulated an iterative, convergent method to approximate the marginal posterior distribution given the joint posterior distribution. The iterative algorithm was inspired by variational Bayesian method and minorization maximization techniques.

Objective and subjective evaluation showed its effectiveness over existing methods in the power spectrum domain, especially in the perceptual quality of the dereverberated audio. Moreover, it showed that the infinite model performs better than an equivalent model formulated as a finite model. We also evaluated our model using audio-to-score alignment task in MIR, and found that dereverberation improves the robustness of score alignment method, suggesting that other MIR tasks may merit from having a dereverberation front-end.

Future work include hierarchical modeling of the presented model, so that human intervention is completely unnecessary. In particular, a model that does require the user to set U_0, V_0 and α is expected to make the system more robust, as these parameters tend to be counter-intuitive for the end-users. Real-time inference is another future work.

[5]LEB01 was omitted because KAM09 consistently outperformed LEB01 in terms of ISD, suggesting that KAM09 produces dereverberated signals that perceptually sounds closer to the dry signal.

[6]We provide sample audio signals at:
`http://winnie.kuis.kyoto-u.ac.jp/members/amaezaw1/taslp2014`

APPENDIX A
PROOF OF (22)

Theorem A.1: For a differentiable and strictly monotonic function $f(x)$, a positive density function q over the domain of x and some $\mathcal{C}$ such that $0 < f(x) < \mathcal{C}$ for all x in its domain, the following holds for all $\mathcal{V} \in (0, \mathcal{C})$:

$$\langle -f(x) \rangle_{q(x)} \geq \mathcal{V} \langle \log(\mathcal{C} - f(x)) \rangle_{q(x)} - \mathcal{V} \log \mathcal{V} - \mathcal{C} + \mathcal{V}. \quad (40)$$

The bound is tight at $\mathcal{V} = \exp \langle \log(\mathcal{C} - f(x)) \rangle_{q(x)}$.

Proof: Consider a function $\bar{u} \log(\mathcal{C} - f(x)) + \bar{v}$ for some $\bar{u} > 0$ and $\bar{v}$. Match the tangent of this function and that of $-f(x)$ about $x = x_0$ such that $f(x_0) \in (0, \mathcal{C})$, and solve for $\bar{u}$ and $\bar{v}$. Set $\mathcal{V} = \mathcal{C} - f(x_0)$ and take the expectation of both sides to obtain (40). Since the r.h.s. is strictly concave, it meets the l.h.s. at exactly one point.

To tighten the bound, optimize the r.h.s. with respect to x_0 to obtain $f(x_0) = M - \exp \langle \log(\mathcal{C} - f(x)) \rangle_{q(x)}$. Substitute this to the definition of $\mathcal{V}$ to tighten the bound w.r.t. $\mathcal{V}$ by setting it to $\exp \langle \log(\mathcal{C} - f(x)) \rangle_{q(x)}$. ∎

Apply this theorem, where $f(\beta, w) = \langle N \rangle_{q(S)}$ and $q(x) = q(\beta, w)$ to obtain the desired inequality in (22).

Next, we shall derive $\mathcal{C}_{\min}$. First note that $\log(\mathcal{C}(f,t) - \langle N(f,t) \rangle_{q(S)q(Z)})$ is defined for the domain of $q(w)q(\beta)$; therefore, for all w' in $\bar{I}$-simplex and $\beta' \in [0,1]$, $\mathcal{C}(f,t) - N(f,t)_{q(S)q(Z)}|_{w=w',\beta=\beta'}$ should be greater than zero. Note that $\langle N(f,t) \rangle_{qS}$ can be expressed an inner product of two vectors $\mathbf{v}(f,t)$ and $\mathbf{w}(f,t)$, where

$$\mathbf{v}(f,t) = \begin{pmatrix} \beta(f) \\ (1-\beta(f) w_1(f) \\ (1-\beta(f) w_2(f) \\ \cdots \\ (1-\beta(f) w_{\bar{I}}(f) \end{pmatrix} \text{ and } \mathbf{w}(f,t) = \begin{pmatrix} \langle S(f,t) \rangle_{q(S(f,t))} \\ Y(f,t-1) \\ Y(f,t-2) \\ \cdots \\ Y(f,t-\bar{I}) \end{pmatrix},$$

with $\sum_i \mathbf{v}_i(f,t) = 1$. Since $|\mathbf{v}(f,t)| \leq 1$, we can bound $|\mathbf{v}(f,t) \cdot \mathbf{w}(f,t)| \leq |\mathbf{w}(f,t)|$. Therefore, the bound is valid if $\mathcal{C}(f,t) > |\mathbf{w}(f,t)|$.

APPENDIX B
LIST OF DISTRIBUTIONS

$$\mathrm{Gam}(X|U,V) \propto X^{U-1} e^{-VX}$$
$$\mathrm{Beta}(X|A,B) \propto X^A (1-X)^B$$
$$\mathrm{Dir}(w|\alpha) \propto \prod_i w_i^{\alpha_i - 1}$$

REFERENCES

[1] A. Tsilfidis and J. Mourjopoulos, "Blind single-channel dereverberation for music post-processing," in *Audio Engineering Society Convention*, 5 2011.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114 –126, Nov. 2012.

[3] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 18, no. 3, pp. 649–662, 2010.

[4] A. Maezawa, H. Okuno, T. Ogata, and M. Goto, "Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden markov model," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, May 2011, pp. 185 –188.

[5] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, 2009, pp. 45–48.

[6] T. Yoshioka, "Speech enhancement in reverberant environments," Ph.D. dissertation, Kyoto University, 2010.

[7] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[8] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[9] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "A sticky HDP-HMM with application to speaker diarization," *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[10] K. Yoshii and M. Goto, "A nonparametric Bayesian multipitch analyzer based on infinite Latent Harmonic Allocation," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 20, no. 3, pp. 717–730, Mar. 2012.

[11] A. Sehr, R. Maas, and W. Kellermann, "Frame-wise HMM adaptation using state-dependent reverberation estimates," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, 2011, pp. 5484–5487.

[12] R. Mukai, S. Araki, and S. Makino, "Separation and dereverberation performance of frequency domain blind source separation," in *Proc. Eurospeech*, 2001, pp. 2599–2602.

[13] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145 –152, Feb. 1988.

[14] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Intl. Workshop on Acoust. Echo and Noise Control*, 2003, pp. 99–102.

[15] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 17, no. 4, pp. 534–545, 2009.

[16] T. Nakatani, M. Miyoshi, and K. Kinoshita, "One microphone blind dereverberation based on quasi-periodicity of speech signals," in *Adv. Neural Info. Process. Syst.* MIT Press, 2003, pp. 1417–1424.

[17] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, 2003, pp. 92–95.

[18] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 1, pp. 69–84, 2011.

[19] T. Yoshioka, H. Kameoka, T. Nakatani, and H. G. Okuno, "Statistical models for speech dereverberation," in *Proc. Workshop on Applications of Signal Process. to Audio and Acoust.*, 2009, pp. 145–148.

[20] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 18, no. 7, pp. 1717 –1731, Sept. 2010.

[21] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Adv. Neural Info. Process. Syst.* MIT Press, 2001.

[22] C. Evers and J. Hopgood, "Multichannel online blind speech dereverberation with marginalization of static observation parameters in a Rao-Blackwellized particle filter," *J. Signal Process. Syst.*, vol. 63, no. 3, pp. 315–332, 2011.

[23] N. Yasuraoka, T. Yoshioka, T. Nakatani, A. Nakamura, and H. G. Okuno, "Music dereverberation using harmonic structure source model and wiener filter," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, 2010, pp. 53–56.

[24] K. Kumar, B. Raj, R. Singh, and R. Stern, "An iterative least-squares technique for dereverberation," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, May 2011, pp. 5488 –5491.

[25] J. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 18, no. 7, pp. 1746 –1765, Sept. 2010.

[26] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[27] K. Lebart and J. Boucher, "A new method based on spectral subtraction for speech dereverberation," *ACUSTICA*, vol. 87, no. 3, pp. 359–366, May-Jun 2001.

[28] A. Abramson, E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation using GARCH modeling," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, 2008, pp. 4565–4568.

[29] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, University College London, 2003.

[30] N. Bryan and G. Mysore, "An efficient posterior regularized latent variable model for interactive sound source separation," in *Proc. Intl. Conf. on Machine Learning*, 2013, pp. 208–216.

[31] O. Dikmen and C. Fevotte, "Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5163–5175, 2012.

[32] W. Ewens, "Population genetics theory - the past and the future," in *Mathematical and Statistical Developments of Evolutionary Theory*, S. Lessard, Ed. Kluwer Academic Publishers, 1990, pp. 177–227.

[33] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1205–1215, 2011.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[35] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *Proc. Intl. Conf. on Machine Learning*. New York, NY, USA: ACM, 2006, pp. 969–976.

[36] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American statistician*, pp. 30–37, 2004.

[37] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Adv. Neural Info. Process. Syst.*, 2000, pp. 556–562.

[38] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition." in *Proc. Intl. Conf. on Language Resources and Evaluation*. European Language Resources Association, 2000.

[39] J. Merimaa, T. Peltonen, and T. Lokki, "Concert hall impulse responses - Pori, Finland: Reference," Retrieved from http://www.acoustics.hut.fi/projects/poririrs on 1/27/2003, 2005.

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

[41] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[42] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin: Springer, 2008.

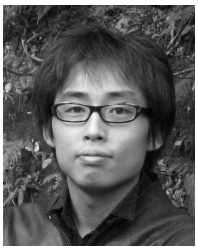
[43] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling," in *Proc. Intl. Conf. on Acoust., Speech and Signal Process.*, May 2011, pp. 3816 –3819.

[44] A. Arzt, G. Widmer, and S. Dixon, "Automatic page turning for musicians via real-time machine listening," in *Proc. 18th European Conf. on Artificial Intelligence*, 2008.

[45] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. Intl. Conf. on Music Info. Retrieval*, 2002, pp. 287–288.

Akira Maezawa received his B.S. degree in electrical engineering (*summa cum laude*) from State University of New York at Binghamton in 2008 and M.S. degree in informatics from Kyoto University, Kyoto, Japan in 2011. He has joined Yamaha Corporation since 2011. He is also currently pursuing a Ph.D degree at Kyoto University since 2013. His research interests include music synchronization, music performance analysis and application of Bayesian inference to statistical signal processing. He is a member of the IPSJ and ASJ.

Katsutoshi Itoyama (M'13) received the B.E. degree in 2006, the M.S. degree in Informatics in 2008, and the Ph.D. degree in Informatics in 2011 all from Kyoto University. He is currently an Assistant Professor of the Graduate School of Informatics, Kyoto University, Japan. His research interests include musical sound source separation, music listening interfaces, and music information retrieval. He recieved the 24th TAF Telecom Student Technology Award and the IPSJ Digital Courier Funai Young Researcher Encouragement Award. He is a member of the IPSJ, ASJ, and IEEE.



Kazuyoshi Yoshii received the Ph.D degree in Informatics in 2008 from Kyoto University, Japan. He is currently a Senior Lecturer at Kyoto University. He has received several awards including the IPSJ Yamashita SIG Research Award and the Best-in-Class Award of MIREX 2005. His research interests include music signal processing and machine learning. He is a member of the Information Processing Society of Japan (IPSJ) and Institute of Electronics, Information and Communication Engineers (IEICE).

Hiroshi G. Okuno (F'12) received B.A. and Ph.D from the University of Tokyo in 1972 and 1996, respectively. He worked for NTT, JST, Tokyo University of Science, and Kyoto University. He is currently a professor of Graduate Program for Embodiment Informatics, Graduate School of Creative Science and Engineering, Waseda University, and a professor emeritus, Kyoto University. He was visiting scholar at Stanford University from 1986 to 1988. He is engaged in computational auditory scene analysis, music information processing and robot audition. He received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001, 2005, 2010, and 2013, IEEE/RSJ IROS-2001 and 2006 Best Paper Nomination Finalist, and NTF Award for Entertainment Robots and Systems in 2010. He co-edited "Computational Auditory Scene Analysis" (Lawrence Erlbaum Associates, 1998), "Advanced Lisp Technology" (Taylor and Francis, 2002), and "New Trends in Applied Artificial Intelligence (IEA/AIE)" (Springer, 2007). He is a fellow of the Japanese Society for Artificial Intelligence, and a member of AAAI, ACM, ASA, RSJ, IPSJ, JSSST and JCSST.