Modal decomposition of musical instrument sounds via optimization-based non-linear filtering

Yoshiki Masuyama*, Tsubasa Kusano†, Kohei Yatabe‡ and Yasuhiro Oikawa§

Department of Intermedia Art and Science, Waseda University, Ohkubo 3–4–1, Shinjuku-ku, Tokyo, 169–8555 Japan

(Received 1 March 2018, Accepted for publication 7 December 2018)

Abstract: For musical instrument sounds containing partials, which are referred to as modes, the decaying processes of the modes significantly affect the timbre of musical instruments and characterize the sounds. However, their accurate decomposition around the onset is not an easy task, especially when the sounds have sharp onsets and contain the non-modal percussive components such as the attack. This is because the sharp onsets of modes comprise peaky but broad spectra, which makes it difficult to get rid of the attack component. In this paper, an optimization-based method of modal decomposition is proposed to overcome it. The proposed method is formulated as a constrained optimization problem to enforce the perfect reconstruction property which is important for accurate decomposition and causality of modes. Three numerical simulations and application to the real piano sounds confirm the performance of the proposed method.

Keywords: Convex optimization, Alternating direction method of multipliers (ADMM), Perfect reconstruction, Causality, Pianos

PACS number: 43.75.Zz [doi:10.1250/ast.40.186]

1. INTRODUCTION

Modal decomposition is one of the most fundamental tools for analyzing musical instrument sounds containing partials, which are referred to as modes, because the decaying processes of the modes are essential factors of the timbre of musical instruments [1,2]. Each mode may decay with a complicated decay process, which characterizes the sound. For example, considering the piano, transfer of energy among coupled strings, bridge, and soundboard causes special decay patterns of the modes, such as double decay and beats, which make the piano sound distinctive [3–7]. Since modes contain such significant information on musical instrument sounds, modal decomposition plays an important role in the analysis of musical instrument sounds, and several modal analysis methods have been proposed [8,9].

Modal decomposition also plays an important role in synthesizing musical instrument sounds based on the models described later. As modes contain significant information on the corresponding sound, parametric modeling of each mode is often considered in the context of sound synthesis. Many models have been proposed in this respect, including exponentially damped sinusoidal (EDS) model [10–14] and damped and delayed sinusoidal model [15–18] as an extension of EDS model. To represent complex decaying processes such as double decay and beats, recently, adaptive harmonic model (AHM) has been applied to the modeling of musical instrument sounds [19–22]. In AHM, each mode is represented by the product of time-varying amplitude and a frequency modulated sinusoid. For using these models, an accurate modal decomposition method is required for estimating model parameters, especially in AHM.

In general, a signal without strong frequency modulation is decomposed by linear filtering. In modal decomposition of musical instrument sounds, a filterbank constructed by the linear bandpass filters is utilized [23]. However, linear filtering causes phase delay and/or pre-ringing as shown in Fig. 1. These are not eliminated simultaneously because of the trade-off between the causality and zero-phase response. Meanwhile, the short-time Fourier transform (STFT) based methods have also been utilized [7], which are suitable for obtaining the outline of each mode in the decay part. However, STFT can be interpreted as a filterbank, and hence STFT inherits the issues of linear filtering.
Gentle onset optimization-based method of modal decomposition is without phase delay and pre-ringing. Hence, the spectra, and modal analysis around the onset is often have sharp onsets is difficult because of such broadness of the spectrum of the filtered signal is narrowband corrupted by pre-ringing as illustrated in Fig. 2(b). This is in contrast, after bandpass filtering, its sharp onset is damped sinusoid) whose waveform has a sharp onset. The spectrum of such modal signal (imitated by an exponentially damped sinusoid) whose waveform has a sharp onset. In contrast, after bandpass filtering, its sharp onset is corrupted by pre-ringing as illustrated in Fig. 2(b). This is because the spectrum of the filtered signal is narrowband as shown in Fig. 2(b). Therefore, extracting modes which have sharp onsets is difficult because of such broadness of the spectra, and modal analysis around the onset is often ignored.

This study aims to realize accurate modal decomposition without phase delay and pre-ringing. Hence, the optimization-based method of modal decomposition is proposed to overcome the accuracy deficiency around the onset \(^{[24]}\). The proposed modal decomposition is formulated as the minimization of the energy of the modes weighted in the frequency domain with the perfect reconstruction and causality constraints. The important differences between the proposed method and the other methods mentioned above are three-fold: (1) the attack component is handled in modal decomposition explicitly; (2) the phase delay and pre-ringing are eliminated simultaneously by constraints; and (3) the fidelity to the data is considered in the perfect reconstruction constraint only. Consequently, the proposed modal decomposition can be considered as non-linear filtering. For optimization, the alternating direction method of multipliers (ADMM) is utilized, which enables the rapid acquisition of results. Moreover, the procedures in ADMM can be calculated in closed form, which results in low computational cost per iteration. The utility of the proposed modal decomposition is demonstrated through three numerical simulations and application to the real piano sound.

2. PRELIMINARIES

A mode of musical instrument sounds is a component corresponding to a single spectral peak. In this paper, a sound which comprises an attack followed by decaying partials without noticeable frequency modulation is considered, such as percussive, plucked string, or struck string instruments. That is, signals with strong frequency modulation, such as vibrato, are beyond the scope of this paper. Based on this assumption, modal decomposition using a filterbank is considered here. Modal decomposition based on a filterbank is commonly utilized, but there are some potential issues because it is based on linear filtering.

2.1. Potential Issues of Linear Filtering

It is well known that linear filtering cannot achieve causality without phase delay as illustrated in Fig. 1. If a filter is causal, i.e., no component is generated before the onset of the original musical instrument sounds, there exists some phase delay which shifts the waveform. Considering an impulse response of causal filters \( h^{\text{causal}}(t) = 0 \) for all time \( t < 0 \), it can be written as

\[
 h^{\text{causal}}(t) = h^{\text{even}}(t) + \text{sign}(t)h^{\text{odd}}(t),
\]

\(^{1}\)The preliminary version of this paper has appeared in conference proceedings \([24]\), where the difference between \([24]\) and this paper is as follows. In this paper, we show the theoretical limitation of conventional linear filtering in Sect. 2. The proposed method is explained in more detail in Sect. 3, especially the weight design which contains an improvement. Furthermore, the relations between the hyperparameters of the proposed method and its performance are discussed in the numerical experiments in Sect. 4. The number of experiments is also increased to investigate the property of the proposed method.
where $h_{\text{even}}(t)$ is the impulse response given by an even function. In general, if $h_{\text{even}}(t)$ is a real and even function, the corresponding frequency response $\tilde{h}_{\text{even}}(\omega)$ is also a real and even function, where $\omega$ is the angular frequency. Then, $\tilde{h}_{\text{causal}}(\omega)$ can be written as

$$\tilde{h}_{\text{causal}}(\omega) = \tilde{h}_{\text{even}}(\omega) - \frac{j}{\omega} * \tilde{h}_{\text{even}}(\omega),$$

(2)

$$= \left(1 - \frac{j}{\omega}\right) * \tilde{h}_{\text{even}}(\omega),$$

(3)

where $j = \sqrt{-1}$, and * is the convolution. Equation (3) shows that causal filters must cause phase delay owing to the convolution of $(1 - j/\omega)$. Such phase delay corrupts the accuracy of modal decomposition.

On the other hand, if a filter does not have phase delay (zero-phase), then the components so-called pre-ringing exist before the onset. In this case, the frequency response is given by a real and even function $\tilde{h}_{\text{even}}(j\omega)$. Then, the impulse response should be represented by a real and even function $h_{\text{even}}(t)$, which indicates that the impulse response exists before the onset in the time domain as the pre-ringing. This trade-off between the phase delay and pre-ringing indicates that a linear filter cannot avoid deformation of the waveform of an extracted mode. The steep attack of modes is always corrupted by phase delay and/or pre-ringing caused by linear filtering, and thus modal decomposition accurate around the onset cannot be accomplished by linear filtering. In other words, an accurate decomposition method must be a non-linear process.

### 2.2. Interpretation of Linear Filtering as a Least Squares Method

Let a discrete signal of given musical instrument sounds be denoted by $s$ whose $r$th element is $s_r$ with the time index $r = 1, \ldots, L$, where $L$ is the length of the signal. Then, its Fourier spectrum is represented by $\hat{s}$ whose $\xi$th element is $\hat{s}_\xi$ with the frequency index $\xi = 1, \ldots, L$. Hereafter, the Fourier transform of $z$ is denoted by $\hat{z}$ ($=Fz$), where $F \in \mathbb{C}^{L \times L}$ is the Fourier transform matrix. Assuming the above condition to the signal $s$, the $k$th mode can be extracted by linear filtering in the frequency domain as

$$\hat{x}_k = \hat{H}_k \hat{s},$$

(4)

where $\hat{H}_k \in \mathbb{C}^{L \times L}$ is a diagonal matrix whose diagonal elements are the frequency response of a predefined filter $\hat{h}_k \in \mathbb{C}^L$ designed specifically for the $k$th mode $x_k$. By preparing $K$ narrowband filters corresponding to the $K$ modes, linear filtering given by Eq. (4) approximately obtains the modes. The accuracy of this decomposition depends on the design strategy of $\hat{h}_k$.

Here, an interpretation of linear filtering as a least squares method is introduced, which will illustrate the relation between the proposed method and the ordinary linear filtering. Let a filter $\hat{H}_k$ admit the inverse $\hat{H}_k^{-1}$. Then, the linear filtering in Eq. (4) can be rewritten as

$$\hat{H}_k^{-1} \hat{x}_k = \hat{s},$$

(5)

which can be interpreted as a least squares method,

$$\min_{\hat{x}_k} \frac{1}{2} \|\hat{H}_k^{-1} \hat{x}_k - \hat{s}\|^2_2,$$

(6)

where $\| \cdot \|_2$ is the Euclidean norm, and its solution $\hat{x}_k$ coincides with the original filtering given by Eq. (4). This interpretation indicates that a linear filtering can be recast to an optimization problem which is a more flexible form as one can modify the formulation easily.

To consider all modes simultaneously, modal decomposition by a filterbank can be considered as illustrated in Fig. 3. Let all modes be denoted by $x$, then Eq. (6) for every $k$ is combined into a single least squares method,

$$\min_{\hat{x}} \frac{1}{2} \|\hat{H}^{-1} \hat{x} - \hat{d}\|^2_2,$$

(7)

where

$$\hat{x} = [\hat{x}_1^T, \ldots, \hat{x}_K^T]^T \in \mathbb{C}^{KL},$$

(8)

$$\hat{d} = [\hat{d}_1^T, \ldots, \hat{d}_K^T]^T \in \mathbb{C}^{KL},$$

(9)

and $\hat{H} \in \mathbb{C}^{KL \times KL}$ is the diagonal matrix whose diagonal elements are given by $[\hat{h}_1^T, \ldots, \hat{h}_K^T]^T$, where $\hat{h}_k^T$ is the transpose of $\tilde{h}_k$. This representation allows a compact notation of the proposed method in the next section.

### 3. PROPOSED METHOD

As in the previous section, linear filtering can be interpreted as a least squares method. This point of view allows us to incorporate additional constraints into the filtering process. In this section, we propose a modal decomposition method by adding constraints into the least squares method so that the undesirable trade-off discussed in Sect. 2.1. is avoided, which results in higher accuracy compared to the linear filterbank, especially around the onset. The proposed method comprises two constraints, perfect reconstruction and causality, where each constraint is explained one-by-one in the preceding subsections. In the proposed method, weights are utilized instead of a filterbank and designed using auto-regressive (AR) approximation of the signal. The block diagram of the proposed method is shown in Fig. 4.
3.1. Constraint of Perfect Reconstruction Condition
For accurate modal decomposition, a perfect reconstruction property is considered first. We say that the decomposed modes satisfy the perfect reconstruction condition when

\[ s = \sum_{k=1}^{K} \mathbf{x}_k, \]  

holds. That is, the decomposed modes can reconstruct the original signal perfectly by simply adding them all. Since the Fourier transform matrix \( \mathbf{F} \) is a unitary matrix, \( \text{Eq. (10)} \) is equivalent to

\[ \hat{s} = \sum_{k=1}^{K} \hat{\mathbf{x}}_k. \]  

By imposing this property into \( \text{Eq. (7)} \), an optimization problem of modal decomposition with a perfect reconstruction constraint is defined as

\[ \min_{\hat{s}} \frac{1}{2} \| \mathbf{H}^{-1} \hat{s} - \mathbf{d} \|^2 \quad \text{s.t.} \quad \hat{s} = \sum_{k=1}^{K} \hat{\mathbf{x}}_k. \]  

After solving this problem, modes satisfying the perfect reconstruction condition can be obtained.

However, Eq. (12) cannot handle the non-modal component around the onset. To address this issue and other problems, the formulation is modified in three parts. First, let \( \mathbf{H}^{-1} \) be replaced by an arbitrary diagonal matrix \( \mathbf{W} \). This modification allows a zero in the diagonal entries of \( \mathbf{W} \), while \( \mathbf{H}^{-1} \) does not allow it owing to the inversion. Owing to this modification, \( \mathbf{W} \) can include 0 in its diagonal element, and then the decomposed modes can be more exclusive of each other. Second, to handle the non-modal component such as the attack component, let a residual \( \mathbf{x}_{K+1} \in \mathbb{R}^L \), which is expected to be a pulse at the onset index, also be considered. The perfect reconstruction condition is relaxed to \( s = \sum_{k=1}^{K+1} \mathbf{x}_k \), where a non-modal component is allowed in \( \mathbf{x}_{K+1} \). Third, since the fidelity to the data is considered in both the first and second terms of Eq. (12), the data in the first term are omitted (the relation between the data and decomposed components is considered only in the constraint).

Based on these three modifications, a modal decomposition problem of the following form in the frequency domain is considered:

\[ \min_{\hat{\mathbf{x}}} \frac{1}{2} \| \mathbf{W} \hat{\mathbf{x}} \|^2 \quad \text{s.t.} \quad \hat{\mathbf{x}} = \sum_{k=1}^{K+1} \hat{\mathbf{x}}_k, \]  

where \( \hat{\mathbf{x}} = [\hat{x}_1^T, \ldots, \hat{x}_{K+1}^T]^T \in \mathbb{C}^{(K+1)L} \), the diagonal matrix \( \mathbf{W} \in \mathbb{C}^{(K+1)L \times (K+1)L} \) is given by setting \([w_1^T, \ldots, w_{K+1}^T]^T\) to its diagonal elements, and \( w_k \in \mathbb{C}^L \) is a given weight in the frequency domain. The weight design is described in Sect. 3.5.

3.2. Closed Form Solution to Eq. (13)
Let \( \mathbf{W}_k \in \mathbb{C}^{L \times L} \) be the diagonal matrix whose diagonal elements are given by \( w_k \). Equation (13) can be rewritten into an unconstrained optimization problem,

\[ \min_{\{\mathbf{x}_k\}_{k=1}^{K}} \frac{1}{2} \sum_{k=1}^{K} \| \mathbf{W} \mathbf{x}_k \|^2 + \frac{1}{2} \| \mathbf{W}_{K+1} \left( \hat{\mathbf{s}} - \sum_{k=1}^{K} \mathbf{x}_k \right) \|^2, \]  

which can be solved for each frequency separately:

\[ \min_{\{\mathbf{x}_k, \hat{\mathbf{s}}\}_{k=1}^{K}} \frac{1}{2} \sum_{k=1}^{K} \| w_{k,\xi} \hat{\mathbf{x}}_{k,\xi} \|^2 + \frac{1}{2} \| w_{(K+1),\xi} \left( \hat{\mathbf{s}}_{\xi} - \sum_{k=1}^{K} \mathbf{x}_{k,\xi} \right) \|^2, \]  

where \( \xi \) is the frequency index, \( w_{k,\xi} \) is the \( k \)th element of \( \mathbf{w}_k \), \( \hat{\mathbf{x}}_{k,\xi} \) is the \( \xi \)th element of \( \hat{\mathbf{x}}_k \), and \( \hat{\mathbf{s}}_{\xi} \) is the \( \xi \)th element of \( \hat{\mathbf{s}} \). The solution to Eq. (15) is obtained by

\[ \hat{x}_{k,\xi} = \frac{\prod_{j \neq \xi} |w_{j,\xi}|^2}{\sum_{i=1}^{K} \prod_{j \neq \xi} |w_{j,\xi}|^2} \hat{s}_{\xi} = g_{k,\xi} \hat{s}_{\xi}, \]  

where \( g_{k,\xi} \in [0, 1] \) is the gain, if the denominator is not zero. Then, the modal decomposition defined by Eq. (13) is given as

\[ \hat{\mathbf{x}}_k = \mathbf{G}_k \hat{\mathbf{s}}, \]  

where \( \mathbf{G}_k \in \mathbb{R}^{L \times L} \) is the diagonal matrix whose diagonal entries are \( g_{k,\xi} \), and \( \hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_L]^T \in \mathbb{R}^L \) is the gain for extracting the \( k \)th mode.

These gains depend on the ratio of the weights. They have a gentle peak at the center frequency of the target mode and sharp dips at that of the other modes. The extracted modes are exclusive of each other at the frequency where \( w_{k,\xi} = 0 \). It can be confirmed from the fact that \( g_{k,\xi} = 1 \) and \( g_{(j \neq k),\xi} = 0 \) if \( w_{k,\xi} = 0 \) for any frequency index \( \xi \). Equation (13) can be interpreted as a zero-phase filterbank with the perfect reconstruction property. This closed form solution to Eq. (13) gives an efficient method to calculate the solution of the proposed modal decomposition via ADMM.

3.3. Proposed Formulation with Causality Constraint
Although the perfect reconstruction property is indispensable for accurate decomposition, it does not eliminate the pre-ringing which deteriorates the accuracy around the onset. Therefore, an additional constraint corresponding to causality is considered:
min \left\{ \frac{1}{2} \|W\hat{x}\|_2^2 \right\} \text{s.t.} \ \left\{ \begin{array}{l} \hat{s} = \sum_{k=1}^{K+1} \hat{x}_k, \\
 \left[ F^{-1} \hat{x}_k \right]_r = 0 \ (r < \tau_k) \end{array} \right. \tag{18}
\]
where \(\tau_k\) is the time index of the onset which is supposed to be estimated from the signal beforehand. Since this causality constraint explicitly eliminates the pre-ringing, modal decomposition without phase delay and pre-ringing is realized by solving Eq. (18).

### 3.4. ADMM Algorithm for Solving Eq. (18)

In this paper, ADMM [25,26] is adopted for solving Eq. (18). ADMM is an algorithm which can solve the following convex optimization problem:

\[
\min_{x \in \mathbb{C}^t \times \mathbb{C}^t} f(x) + g(z) \text{ s.t.} \ x = z, \tag{19}
\]

where \(f\) and \(g\) are proper and lower-semicontinuous convex functions. For any initial values \(z_0, u_0, \) and \(\rho > 0,\) ADMM is given by

\[
\begin{align*}
   x^{[i+1]} &= \text{prox}_{\rho f}(z^{[i]} - u^{[i]}), \\
   z^{[i+1]} &= \text{prox}_{g}(x^{[i+1]} + u^{[i]}), \\
   u^{[i+1]} &= u^{[i]} + x^{[i+1]} - z^{[i+1]},
\end{align*}
\]

where \(i\) is the iteration index, and \(\text{prox}_{f}(\cdot)\) is the proximity operator of \(f\) defined by [27]

\[
   \text{prox}_{f}(y) = \arg \min_{x} f(x) + \frac{1}{2\rho} \|y - x\|_2^2. \tag{23}
\]

For applying the ADMM algorithm to the proposed method in Eq. (18), it is reformulated as the following equivalent problem:

\[
\min_{\hat{s}, \hat{z}} \frac{1}{2} \|W\hat{x}\|_2^2 + \chi_{C_1}(\hat{s}) + \chi_{C_1}(\hat{z}) \text{ s.t.} \ \hat{s} = \hat{z}, \tag{24}
\]

where \(\chi_{C}\) is the indicator function of a closed nonempty convex set \(C,\)

\[
\chi_{C}(x) = \begin{cases} 0 & x \in C \\ \infty & \text{otherwise} \end{cases}, \tag{25}
\]

\(C_1\) and \(C_2\) are the sets corresponding to each constraint in Eq. (18),

\[
C_1 = \left\{ \hat{s} \in \mathbb{C}^{(K+1)L} \ | \ \hat{s} = \sum_{k=1}^{K+1} \hat{x}_k \right\}, \tag{26}
\]

\[
C_2 = \left\{ \hat{z} \in \mathbb{C}^{(K+1)L} \ | \ [F^{-1} \hat{z}]_r = 0 \ (r < \tau_k) \right\}. \tag{27}
\]

Then, by regarding the functions in Eq. (24) as

\[
f(\hat{s}) = \frac{1}{2} \|W\hat{x}\|_2^2 + \chi_{C_1}(\hat{s}), \tag{28}
\]

\[
g(\hat{z}) = \chi_{C_1}(\hat{z}),
\]

the ADMM algorithm for Eq. (18) is obtained as follows:

\[
\hat{s}^{[i+1]} = \text{prox}_{\rho f}(\hat{z}^{[i]} - u^{[i]}), \tag{29}
\]

\[
\hat{z}^{[i+1]} = P_{C_1}(\hat{s}^{[i+1]} + u^{[i]}), \tag{30}
\]

\[
u^{[i+1]} = u^{[i]} + \hat{s}^{[i+1]} - \hat{z}^{[i+1]}, \tag{31}
\]

where \(P_{C_1}\) is metric projection onto \(C_2,\) which can be calculated by

\[
P_{C_1}(z_{\tau}) = \left\{ \begin{array}{ll} 0 & \tau < \tau_k \\
   z_{\tau} & \text{otherwise} \end{array} \right., \tag{33}
\]

and \(C_2 = \{ z \in \mathbb{R}^{(N+1)L} \ | \ z_r = 0 \ (r < \tau_k) \}.
\]

The proximity operator in Eq. (29) for \(\hat{s}\)-update is calculated by

\[
\min_{\hat{s}} \frac{1}{2} \|W\hat{x}\|_2^2 + \frac{1}{2\rho} \|\hat{s} - \hat{y}^{[i]}\|_2^2 \text{s.t.} \ \hat{s} = \sum_{k=1}^{K+1} \hat{x}_k, \tag{34}
\]

where \(\hat{y}^{[i]} = \hat{z}^{[i]} - \hat{u}^{[i]}\). Equation (34) is a constrained optimization problem similar to Eq. (13). Therefore, Eq. (34) can be solved analytically in the same way as Eq. (16), which is written as

\[
\hat{x}_{\hat{k} \hat{\xi}} = v_{\hat{k} \hat{\xi}} + \rho (\mu_{\hat{k} \hat{\xi}} - \eta_{\hat{k} \hat{\xi}}), \tag{35}
\]

where each term is defined as

\[
v_{\hat{k} \hat{\xi}} = \prod_{j \neq \hat{k}} (\|w_{\hat{j} \hat{\xi}}\|^2 + \rho) \delta_{\hat{j} \hat{\xi}}, \tag{36}
\]

\[
\mu_{\hat{k} \hat{\xi}} = \left( \sum_{j \neq \hat{k}} \prod_{j \neq \hat{k}} (\|w_{\hat{j} \hat{\xi}}\|^2 + \rho) \right) \delta_{\hat{j} \hat{\xi}}, \tag{37}
\]

\[
\eta_{\hat{k} \hat{\xi}} = \sum_{j \neq \hat{k}} \prod_{j \neq \hat{k}} (\|w_{\hat{j} \hat{\xi}}\|^2 + \rho) \delta_{\hat{j} \hat{\xi}}, \tag{38}
\]

\[
\zeta_{\hat{k}} = \sum_{\hat{j}=1}^{K+1} \prod_{\hat{j} \neq \hat{k}} (\|w_{\hat{j} \hat{\xi}}\|^2 + \rho). \tag{39}
\]

By substituting \(\hat{y}^{[i]} = \hat{z}^{[i]} - \hat{u}^{[i]}\) in the above formulas, \(\hat{s}\)-update in Eq. (29) can be calculated easily.

Since Eq. (18) is convex, the decomposed result obtained by iterating Eqs. (29)–(31) from arbitrary initial values is globally optimal. While an arbitrary choice is allowable, one preferred choice for the initial value \(z_0\) is a solution to Eq. (13) which is given in Eq. (16). The \(\hat{s}\)-update given by Eq. (34) corresponds to the extension of linear filtering with the perfect reconstruction constraint given by Eq. (13), and a solution to Eq. (13) is not modified by the \(\hat{s}\)-update. Therefore, a solution to Eq. (13) should be closer to the globally optimal solution than a randomly chosen initial value.

### 3.5. Weighting Rule for the Proposed Method

The choice of the weight in Eq. (18) is important because it determines the decomposed result. As the weight
penalizes the energy of each frequency component, the kth mode is dominated by some frequencies at which the weight \( w_k \) contains small values. Conversely, frequencies with large weights do not remain in the result much. That is, the weight should be set small around the center frequency of the target mode and large around that of the non-target modes.

To design such weight, the AR model is utilized. By calculating the AR spectrum \( \hat{s} \) of a signal \( s \), the location of poles \( p \) can be obtained. Since only the poles corresponding to the peaks of the spectrum are important, they should be automatically determined. In this paper, we propose to select the poles by thresholding their magnitude. As a pole closer to the unit circle corresponds to the modal behavior, the poles satisfying \( |p| \approx 1 \) are important for modal decomposition. Therefore, after calculating an AR spectrum of a sufficiently large order, the obtained poles \( p \) whose magnitudes are greater than a predefined threshold are selected. Then, the center frequency \( f_k \) of the kth mode is obtained as the complex argument of the selected pole \( p_k \). The amplitude of the AR spectrum at the center frequency \( a_k = \beta |p_k|/|p_k| \) is also utilized in the weight design in Eq. (42).

Based on the information obtained by AR modeling, a weighting rule for modal decomposition is proposed. First, a resonance filter \( \hat{h}_k \) with a conjugate pair of poles, \( p_k \) and \( \overline{p}_k \), is constructed:

\[
\hat{h}_k(z) = \frac{1}{(z - p_k)(z - \overline{p}_k)}. \tag{40}
\]

Then, it is normalized to have the unit amplitude \( \hat{h}_k = \hat{h}_k/\max_0 |\hat{h}_k(e^{\theta})| \), where \( \theta \in [0, 2\pi) \).

Utilizing these elements, the weighting matrix corresponding to the kth mode is designed by

\[
W_k = W_k^{\text{dip}}W_k^{\text{peaks}}, \tag{41}
\]

where each matrix \( W_k^{(i)} \in \mathbb{C}^{L \times L} \) is a diagonal matrix of \( w_k^{(i)} \in \mathbb{C}^L \) whose elements are calculated as

\[
w_k^{\text{dip},i} = \frac{1}{a_k} \max_0 \left( \frac{1}{|\hat{h}_{k,i}|} - \mu, 0 \right), \tag{42}
\]

\[
w_k^{\text{peaks},i} = \sum_{i \neq k} |\hat{h}_{i,i}|, \tag{43}
\]

where \( \mu > 0 \) is a parameter for adjusting the passband. If \( \mu = 0 \), the weight \( w_k^{\text{dip},i} \) is inversely proportional to \( |\hat{h}_{k,i}| \). If \( \mu = 1 \), \( w_k^{\text{dip},i} \) becomes 0 at the center frequency of the kth mode because \( |\hat{h}_{k,i}| = 1 \) for such frequency, which results in the unit gain in Eq. (17). In this case, the gain becomes \( g_{i \neq k,i} = 0 \) at the other center frequencies, and the extracted modes are exclusive of each other. Furthermore, if \( \mu > 1 \), the weight \( w_k^{\text{dip},i} \) becomes 0 around the center frequency, and the passband of \( g_{i \neq k,i} \) becomes broader as increasing \( \mu \).

From the construction, \( w_k^{\text{dip},i} \) consists of a single dip at the center frequency of the kth mode, and \( w_k^{\text{peaks},i} \) consists of \( K - 1 \) peaks at the other frequencies. An example of the proposed weights is illustrated in Fig. 5. For the residual, a specific weight \( w_{(K+1)}^{\text{allPeaks}} \) is proposed,

\[
w_{(K+1)}^{\text{allPeaks}} = \lambda \sum_{i = 1}^K |\hat{h}_{i,i}|, \tag{44}
\]

which eliminates all modes from the residual, where \( \lambda > 0 \) is a parameter adjusting the energy of the attack component. When the target signal contains much non-modal attack component, \( \lambda \) should be set to a smaller value, and vice versa. Its effect is evaluated by numerical simulations in the next section.

4. NUMERICAL SIMULATION

In this section, to investigate the property of the proposed modal decomposition, the following three points are evaluated by simulations: (1) dependency of accuracy on the tuning parameters; (2) robustness to the onset index estimation error; and (3) robustness to additive noise.

The proposed method was applied to simulated musical instrument sounds synthesized by adding \( K \) simple modes \( s_k \) and the impulse as the attack component \( s_{K+1} \). The simulation conditions are shown in Table 1.

Each mode was synthesized by the impulse responses of resonance filters, whose poles were given by \( p_k = A_k e^{2\pi f_k t}/f_s \), where \( A_k = A_{\text{intercept}} - k A_{\text{slope}} \), \( A_{\text{intercept}} = 0.99985 \), and \( A_{\text{slope}} = 10^{-7} \). The example of the waveform and spectrum of the synthesized signal are displayed in Fig. 6. The parameters \( A_{\text{intercept}} \) and \( A_{\text{slope}} \) were obtained by approximating the piano sound with a 100 order AR model using Burg’s method [29]. Under all conditions, an impulse was added to the simulated modes as the attack component, and its amplitude \( A_{K+1} \) was set to be the same as the maximum absolute value of the 1st mode.
The proposed method was compared with five baseline methods: causal filters, zero-phase filters, and STFT-based methods with three window lengths. The sampling rate, $f_s$, was 20,000 Hz, and the absolute values of the poles of the two types of filters were the same as those of the synthesizing filters. The window and overlap length for STFT were, respectively, 51.2 ms and 25.6 ms in STFT(a), 102.4 ms and 51.2 ms in STFT(b), and 204.8 ms and 102.4 ms in STFT(c). The performance of decomposition was evaluated by the signal-to-distortion ratio (SDR):

$$\text{SDR} = \frac{1}{K} \sum_{k=1}^{K} 20 \log_{10} \frac{\|s_k\|_2}{\|s_k - x_k\|_2},$$

where $s_k$ is the $k$th simulated mode, and $x_k$ is its estimation. We also evaluated SDRs before and after the onset, SDR$_{\text{pre}}$ and SDR$_{\text{post}}$, respectively. That is, the estimated signal was divided at the onset into two parts, $x_{\text{pre},k}$ and $x_{\text{post},k}$, and the denominator of Eq. (45) was modified as $\|x_{\text{pre},k} - s_{\text{pre},k}\|_2$ and $\|x_{\text{post},k} - s_{\text{post},k}\|_2$.

### 4.1. Dependency on Tuning Parameters

In this experiment, we evaluated the dependency on the parameters, $\lambda$, $\mu$, and AR order, of the proposed modal decomposition. The range of $\mu \in [1, \mu_{\text{max}}]$ in Eq. (42) was decided by $\mu_{\text{max}}$, which is the maximum value of $\mu$ such that the zero elements (or passband) of the corresponding weights $w_k (k = 1, \ldots, K)$ do not overlap each other (note that the width of zero elements, or passband, increases as $\mu$ increases). We utilized Eq. (17) as the initial value of the proposed method, and ADMM was iterated 10 times because SDR was not significantly improved beyond.

Figure 7 shows the performances of the proposed modal decomposition for each set of parameters. The parameters, $\lambda$ and $\mu$, varied along the horizontal and vertical axes on the logarithmic scale, respectively. For making the comparison easier, the center of the colorbars (green) was set to the best SDR among the conventional methods. That is, the color above the center of the colorbars represents that the proposed method was better than all conventional methods. Comparing the top and bottom rows, AR order is not so significant as compared to the other parameters. $\mu = \mu_{\text{max}}$ achieved the highest SDR for all $\lambda$ except for Sig. 4 (rightmost column). When $\mu = \mu_{\text{max}}$, the passband of the corresponding gain $g_{k,\xi}$ becomes broader. While the passband of STFT-based methods, which achieved the best performance, is associated with the manually defined window length, that of the proposed method is automatically decided by $\mu$ depending on the signal.

The same figures for SDR$_{\text{pre}}$ and SDR$_{\text{post}}$ are illustrated in Fig. 8. Thanks to the causality constraint, SDR$_{\text{pre}}$ of the proposed modal decomposition exceeded 100 dB, which

### Table 1 Conditions of four simulated signals.

<table>
<thead>
<tr>
<th>$f_1$ [Hz]</th>
<th>$\mu_{\text{max}}$ (AR$_{100}$)</th>
<th>$\mu_{\text{max}}$ (AR$_{500}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig. 1</td>
<td>220</td>
<td>148</td>
</tr>
<tr>
<td>Sig. 2</td>
<td>440</td>
<td>421</td>
</tr>
<tr>
<td>Sig. 3</td>
<td>880</td>
<td>819</td>
</tr>
<tr>
<td>Sig. 4</td>
<td>1,760</td>
<td>1,866</td>
</tr>
</tbody>
</table>

![Fig. 6 Waveform and spectrum of the synthesized signal (Sig. 4 in Table 1).](image)

![Fig. 7 SDR of the decomposed modes obtained by the proposed method.](image)
Causal filter

STFT(b)

STFT(c)

Zero-phase filter

STFT(a)

\( \frac{1}{C22} \)

\( \frac{1}{C22} \)

\( f \)

Eq. (17)) was 36.6 dB, while that of the proposed method was higher than that of the other methods. The difference between the proposed method and the other methods was slight (less than about 0.3 dB).

### 4.2. Comparison with Other Methods

SDR, SDR\(_{\text{pre}}\), and SDR\(_{\text{post}}\) for all methods are summarized in Fig. 9, where the AR order was 500, \( \lambda = 1 \), and \( \mu = \mu_{\text{max}} \). As shown in Fig. 9, SDR of the proposed method was higher than that of the other methods. The causal filters resulted in the lowest SDR because of the pre-ringing. The STFT-based methods caused pre-ringing owing to the estimation error of the onset index, yet it still outperformed the other methods for some parameter choices. This result indicates that the amount of the pre-ringing of the proposed method depends on the estimation error of the onset index, while that of the other methods depends on the manually chosen parameters (such as window length in STFT-based methods).

As shown in Fig. 9, SDR of the proposed method was higher than that of the other methods. The causal filters resulted in the lowest SDR because of the pre-ringing. The STFT-based methods caused pre-ringing owing to the estimation error of the onset index, yet it still outperformed the other methods for some parameter choices. This result indicates that the amount of the pre-ringing of the proposed method depends on the estimation error of the onset index, while that of the other methods depends on the manually chosen parameters (such as window length in STFT-based methods).

### 4.3. Robustness to Onset Index Estimation Error

In this experiment, the robustness to the estimation error of the onset index was evaluated. Since the signals in the scope of the proposed method (including percussive, plucked string, and struck string instruments) contain a sharp onset, detecting the onset is easy. However, many of these signals may contain a small noise just before the onset, for example, the piano sound contains small noise related to the movement of the hammer just before the onset. Therefore, the onset index \( \tau_A \) in the proposed method might be better to set before the onset. Here, the estimated onset index \( \tau_{\text{est}} \) was set to 10 and 20 ms before the true onset index, and the performance was evaluated with AR order 500.

Figure 11 shows the performance of the proposed modal decomposition with the estimation error of the onset index, where the color was adjusted as that shown in Fig. 7. According to SDR\(_{\text{pre}}\), the proposed modal decomposition contained pre-ringing owing to the estimation error of the onset index, yet it still outperformed the other methods for some parameter choices. This result indicates that the amount of the pre-ringing of the proposed method depends on the estimation error of the onset index, while that of the other methods depends on the manually chosen parameters (such as window length in STFT-based methods).
4.4. Robustness to Additive Noise

In this experiment, we evaluated the robustness of the proposed modal decomposition to noise. The Gaussian noise was added so that the signal-to-noise ratio (SNR) of the signal became 30 and 60 dB. Figure 12 shows the performance of the proposed modal decomposition in the noisy case, where the color is same as that in the previous figures. According to $\text{SDR}_{\text{post}}$ for the noisier case ($\text{SNR} = 30 \text{ dB}$), $\mu$ should be smaller than $\mu_{\text{max}}$, while $\mu_{\text{max}}$ achieved the highest $\text{SDR}$ in the clean case as in Fig. 7. This is because the passband of the proposed method becomes narrower as $\mu$ decreases, and the passband should be narrowed for reducing the noise in the decomposed modes. Therefore, $\mu$ should be adjusted based on SNR, which can be automated by estimating SNR from the AR spectrum.

5. APPLICATION TO REAL PIANO

The proposed method was applied to the real piano sounds in the dataset of the University of Iowa\textsuperscript{2}. Piano is a struck string instrument whose onset is sharp and contains complicated an attack component. Therefore, piano sounds are suitable for confirming the utility of the proposed method. In the following subsections, extracted modes and the attack component were compared in two situations (same pitch with different dynamics and different pitches with the same dynamics).

The sampling rate was 44,100 Hz, and the parameters of the proposed method were set to $\lambda = 0.01$ and $\mu = \mu_{\text{max}}$. The poles satisfying $0.999 < |p_k| (< 1)$ were automatically selected for constructing the weight. The onset index $\tau_A$ was also automatically estimated by finding the first sample whose magnitude exceeds $10^{\text{SDR}_{\text{pre}}} / 3$ times the maximum absolute value of the waveform ($-60 \text{ dB}$ from the maximum in the time domain). For comparison, an STFT-based method whose window and overlap length were 92.9 ms and 46.4 ms, respectively, was also applied.

5.1. Sounds with Different Dynamics

The extracted modes and attack component of two different dynamics but same pitch (A4) were compared. The waveforms and spectra of the mezzo forte (soft) and forte (loud) sounds are shown in Fig. 13. The waveforms indicate that the 1st modes of the proposed and STFT-based methods nearly coincide, where the difference appeared as pre-ringing. On the other hand, the 2nd modes around the onset are largely different. The 2nd modes of the STFT-based method seem more random than those of the proposed method. This should be because the complicated attack component of the real piano sound, which have broad spectra, remain in the extracted mode. Such effect of the attack component exists in all modes of the STFT-based method, where the 1st mode was less affected.

\footnotesize\textsuperscript{2}“University of Iowa Electronic Music Studios,” http://theremin.music.uiowa.edu/MISpiano.html
owing to the larger magnitude. The same effect can be confirmed from the spectra in the right columns. While the proposed method retained the non-harmonic components (the slopes from around −50 dB to −100 dB in Fig. 13) in the attack component, the STFT-based method removed them from the residual (see lower right panels in Fig. 14). That is, the STFT-based method included the non-harmonic components in the decomposed modes, while the proposed method did not. This is the unique feature of the proposed method.

5.2. Sounds with Different Pitches

The modes and attack component of the same dynamics but different pitches (G♯4 and A♯4) were compared. The waveforms and spectra are shown in Fig. 15.

In Fig. 16, the modes and the attack component extracted from G♯4 and A♯4 sounds are shown. Similar to Fig. 14, the 2nd modes of the STFT-based method are more random than those of the proposed method around the onset, which is also indicated in the spectra of the attack components in the lower right panels. By seeing the attack component, we found some impulsive noises contained in the A♯4 sound, which can be confirmed via the spectrogram in Fig. 17. These spectrograms indicate that the STFT-based method included the impulsive noise into the decomposed modes, while the proposed method distinguished the modes from these non-modal components.

In Fig. 18, envelopes of the obtained 1st modes are shown. According to the figure, dynamics have a little effect on the decay process of the mode, while the pitch greatly affects it. This coincides with an earlier study [7] that suggests the appropriateness of the decomposition.

6. CONCLUSION

In this paper, the modal decomposition method of musical instrument sounds was proposed. By interpreting a filtering process as a least squares method, the proposed method is formulated as a constrained optimization problem. Hence, an undesired trade-off of linear filtering, which corrupts modal decomposition around the onset, is circumvented. The proposed optimization model is solved by the ADMM algorithm, where the closed form solution of the constrained quadratic problem allows an easy and fast update of the variables. The numerical simulation showed the accuracy of the proposed method which achieved higher SDR than the other methods. The simulation also confirmed that the proposed method works correctly even if there exist estimation errors of the onset index and noise. The application to the real piano sound showed that the proposed method reduced randomness around the onset, which makes the analysis easier.

ACKNOWLEDGEMENT

We thank Mr. Koji Oishi, Mr. Yusuke Miyagi, Mr. Ken
providing useful suggestions. Takahashi, as well as their colleagues at KORG INC., for providing useful suggestions.

REFERENCES


