ONSET-INFORMED SOURCE SEPARATION USING NON-NEGATIVE MATRIX FACTORIZATION WITH BINARY MASKS

Yuta Kusaka†, Katsutoshi Itoyama†, Kenji Nishida† and Kazuhiro Nakadai†‡

† Department of Systems and Control Engineering
Tokyo Institute of Technology
Tokyo, Japan
{kusaka, itoyama, nishida}@ra.sc.e.titech.ac.jp

‡ Honda Research Institute Japan Co., Ltd.
Saitama, Japan
nakadai@jp.honda-ri.com

ABSTRACT

This paper describes a new onset-informed source separation method based on non-negative matrix factorization (NMF) with binary masks. Many previous approaches to separate a target instrument sound from polyphonic music have used side-information of the target that is time-consuming to prepare. The proposed method leverages the onsets of the target instrument sound to facilitate separation. Onsets are useful information that users can easily generate by tapping while listening to the target in music. To utilize onsets in NMF-based sound source separation, we introduce binary masks that represent on/off states of the target sound. Binary masks are formulated as Markov chains based on continuity of musical instrument sound. Owing to the binary masks, onsets can be handled as a time frame in which the binary masks change from off to on state. The proposed model is inferred by Gibbs sampling, in which the target sound source can be sampled efficiently by using its onsets. We conducted experiments to separate the target melody instrument from recorded polyphonic music. Separation results showed about 2 to 10 dB improvement in target source to residual noise ratio compared to the polyphonic sound. When some onsets were missed or deviated, the method is still effective for target sound source separation.

1. INTRODUCTION

The sound source separation problem is a challenging task and separation of a specific instrument sound from a polyphonic sound is a subtask of this problem. The extracted instrument sound is useful for users who desire to practice playing the instrument or produce remixes. The separated sound is also applicable to many systems such as music editing [1], creating a Karaoke sound source [2], automatic transcription systems [3], and recognizing instruments [4, 5]. Furthermore, the separated melody line of the target can be applied in music retrieval systems like query-by-melody ones [6, 7]. Thus, separation of a particular sound from a polyphonic sound is considered to be an important topic.

Non-negative matrix factorization (NMF) [8, 9] or independent component analysis (ICA) [10] have been studied as effective methods for monaural source separation. These methods decompose an input signal into a multiple component set of typical parts in the signal, such as the notes of each instrument. Separation of the target instrument sound is performed by isolating the components corresponding to that instrument from the decomposed components set. However, there is a problem that instrument sounds and the decomposed components do not have one-to-one correspondence. For example, when a piano signal is decomposed, in order to know which of the decomposed components correspond to the C4 sound, it is necessary to follow a procedure like listening to the recovered sound. Besides, when the input is a polyphonic musical signal, the set of decomposed components becomes even more complicated. One can easily imagine how difficult it would be to determine the components corresponding to the target instrument among them.

Informed source separation (ISS) is an approach to separate a target instrument sound by using some side-information such as a template sound or music score [11]. In ISS, the separation is assisted by information such as spectral or temporal structure of the target that is provided in advance. ISS is a powerful approach, however, it has limitations when it is impracticable or difficult to prepare the necessary side-information. User-guided approaches [12, 13] using information that can be created by listening to the music have been proposed, but these approaches also demand a lot of time and effort to create information.

In this paper, we propose a new informed source separation method that we call, *onset-informed NMF* (OI-NMF), which can separate a target instrument sound from a polyphonic sound using some of the target instrument onsets as side-information (Figure 1). The onsets of the target instrument is useful information that can be created by a simple task such as tapping when listening to the target polyphonic music. OI-NMF provides a function to utilize the onsets of the target instrument as side-information in NMF. One of the advantages of OI-NMF is that it does not require every onsets, i.e. it is feasible even when some are missing. Therefore, it is expected that the target instrument separation will be easily performed with music for which preparation of side-information is difficult. The contributions of this paper are as follows:

- We extended the existing NMF model so that onsets of a target sound source could be input as side-information. To treat onsets as time frames where the instrument sound switches from off to on, binary masks based on Markov chain overlaying NMF activation were introduced. We derived an inference algorithm for the OI-NMF defined as a stochastic model in a Bayesian manner.
- We implemented OI-NMF and performed experiments and evaluations using a separated target instrument sound from polyphonic sound. A recorded polyphonic dataset and target instrument onsets generated from the f0 annotation in-

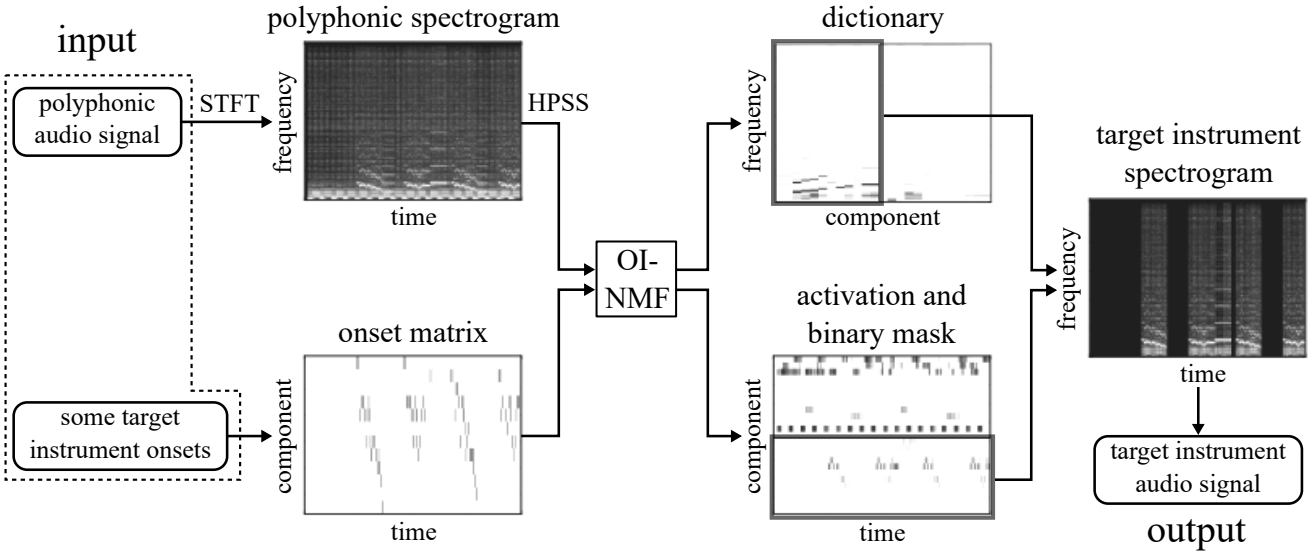


Figure 1: An overview of *onset-informed NMF* (OI-NMF). The input is the polyphonic sound signal and some target instrument onsets. These inputs are converted into a spectrogram and an onset matrix, respectively. The polyphonic spectrogram is decomposed into a dictionary, its activation, and binary mask by OI-NMF. From the components of the results given the onsets, the target instrument spectrogram is isolated.

cluded the dataset were used. The target instrument separation with the onsets was performed at a higher level than the baseline methods without the onsets, and it was confirmed that onsets of a target instrument sound were effective side-information in ISS.

2. RELATED WORKS

OI-NMF can be included in ISS approaches because it performs target instrument separation using its onsets as side-information. The basic ISS approaches are roughly classified based on the type of side-information.

- Spectral structure of the target. If a target is a musical instrument, this corresponds to its tone, timbre, or harmonics. Supervised NMF (SNMF) [14, 15] performs separation based on a pre-learned dictionary from a sound template of the target. In SNMF, if the prepared template does not resemble the target instrument, the separation quality may be degraded. Adding constraints to the model can help improve accuracy.
- Temporal structure of the target. Onsets would be classified in this part. Typical information includes scores of the input music or target instrument. The music score contains the onset and offset of each note indicating when the instrument is active (temporal structure). In addition, the score also contains the pitch of each note (spectral structure). Score-informed NMF [16, 17] achieves separation by using this rich information. Deep learning-based methods, in which learning is performed using a spectrogram of the clean target source as a teacher, are also included in this approach [18, 19].

Side-information described in above methods are effective in instrument sound separation, however, they require preparation. The sound template may change depending on the music, and the score of the input music may not exist. In the deep learning approach, a large dataset for learning is required. Therefore, these methods are sometimes difficult to apply in practice.

To solve this problem, many user-guided ISS methods have been proposed. In these methods, separation is performed by using information that users can create by listening to or observing the multiple instrument music and the target instrument sound. Examples of such user-guided information are humming in mimicking the target [12], or annotation of the target region on the polyphonic spectrogram [13]. Although such side-information can be suitably created to match the music and the target instrument sound, its creation often demands much time and high skill. Compared to these methods, the simplicity of creating onsets has an advantage.

A multichannel non-negative tensor factorization model [20] can be regarded as a similar method to OI-NMF from the viewpoint that temporal information created by the user is utilized to support the separation. This method requires more work to prepare side-information than OI-NMF since it needs offsets as well as onsets.

3. TARGET SOURCE SEPARATION BASED ON NMF

NMF [8, 9] is an algorithm that is effective for sound source separation for a monaural signal. It was originally proposed in image processing [21] and its applicability has been investigated, for example, sound source separation [22, 23], and automatic transcription [24]. In the context of sound source separation, NMF decomposes the amplitude magnitude spectrogram obtained by short-time Fourier transform (STFT) of an input audio signal from multiple instruments based on its low-rank property,

$$\boldsymbol{X} \simeq \boldsymbol{W}\boldsymbol{H} \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{F\times T}$ is a magnitude spectrogram, $f = 1, 2, \ldots, F$ is a frequency bin and $t = 1, 2, \ldots, T$ is a time frame index of $\boldsymbol{X}$. Output matrix $\boldsymbol{W} \in \mathbb{R}^{F\times K}_{\geq 0}$ is a dictionary matrix, in which each component $k = 1, 2, , \ldots, K$ represents the spectral patterns included in $\boldsymbol{X}$. The other matrix $\boldsymbol{H} \in \mathbb{R}^{K\times T}_{\geq 0}$ is an activation ma-

trix which represents the time change of the corresponding spectral patterns in $\boldsymbol{W}$. K is a parameter that determines the number of components and is given in advance.

In order to reconstruct the target instrument source from the NMF results, a set of components corresponding to the target instrument sound is selected, and a target magnitude spectrogram $\boldsymbol{X}_{\text{target}}$ is constructed by using Wiener filter,

$$\boldsymbol{X}_{\text{target}} = \frac{\boldsymbol{W}_{\text{target}}\boldsymbol{H}_{\text{target}}}{\boldsymbol{W}\boldsymbol{H}} \circ \boldsymbol{X}, \tag{2}$$

where $\boldsymbol{W}_{\text{target}}$ and $\boldsymbol{H}_{\text{target}}$ are the dictionary and its activation consisting of only components corresponding to the target, and $\circ$ represents element-wise product. Finally, the target sound source is obtained by performing inverse STFT to $\boldsymbol{X}_{\text{target}}$ and the phase spectrogram. The phase spectrogram of $\boldsymbol{X}$ is sufficient and a estimated phase from $\boldsymbol{X}_{\text{target}}$ may improve the quality of the separated sound [25].

4. ONSET-INFORMED SOURCE SEPARATION

In this section, we describe a new source separation method for music instrument sounds, named onset-informed NMF (OI-NMF), which uses onsets of the target instrument. Figure 1 shows an overview of OI-NMF. The input is the magnitude spectrogram of a multiple instrument sound analyzed by STFT and some onsets of the target instrument sound. The output is a magnitude spectrogram of the target instrument sound. In OI-NMF, the components that correspond to the target instrument sound appear on the components given the onsets. Therefore, the target instrument sound can be obtained by applying Wiener filter (Equation 2) to these components and performing ISTFT.

The main feature of OI-NMF is binary masks introduced for the NMF activation to treat the target instrument onsets as side-information. The binary masks are defined as a binary matrix the same size as the activation, and controls on/off of the activation, that is, on/off of the instrument sound by its 1/0 values. Owing to the binary masks, the onset is considered as a time frame when the binary mask changes from 0 to 1.

Beta process sparse NMF (BP-NMF) [26, 27] is proposed as NMF in which binary masks are introduced. In accordance with BP-NMF, the decomposition of the magnitude spectrogram by OI-NMF is formulated as

$$\boldsymbol{X} \simeq \boldsymbol{W}(\boldsymbol{H} \odot \boldsymbol{S}), \tag{3}$$

where $\boldsymbol{X} \in \mathbb{Z}_{\geq 0}^{F\times T}$ is the input magnitude spectrogram ($\mathbb{Z}_{\geq 0}$ represents the set of all non-negative integers), and $\boldsymbol{W}, \boldsymbol{H}$ are the dictionary and its activation as in Equation (1) respectively. $\boldsymbol{S} \in \{0,1\}^{K\times T}$ is a binary mask matrix of the activation $\boldsymbol{H}$. Note that $\odot$ represents Hadamard product. Prior distributions are set to $\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H}$, and $\boldsymbol{S}$, and the model is treated as a hierarchical Bayesian model. For $\boldsymbol{X}, \boldsymbol{W}$, and $\boldsymbol{H}$, the same priors in BP-NMF are set.

$$X_{f,t} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S} \sim \text{Poisson}\left(X_{f,t} \,\middle|\, \sum_{k=1}^{K} W_{f,k}H_{k,t}S_{k,t}\right), \tag{4}$$

$$W_{f,k} \sim \text{Gamma}\left(W_{f,k} \,\middle|\, \alpha^W, \beta^W\right), \tag{5}$$

$$H_{k,t} \sim \text{Gamma}\left(H_{k,t} \,\middle|\, \alpha^H, \beta^H\right), \tag{6}$$

where $\alpha^W, \beta^W, \alpha^H, \beta^H$ are hyperparameters of gamma distribution. Gamma distribution is represented as

$$\text{Gamma}\left(x \mid \alpha, \beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{7}$$

where $\alpha, \beta > 0$ are hyperparameters. α is called the shape parameter, and when it takes a value less than 1, the value sampled from the gamma distribution tends to be 0. Therefore, α^W is set to smaller than 1 to express the sparsity of the harmonic structure of the instrument sound. On the other hand, if the value of activation becomes 0, the binary mask will not work. Therefore, α^H is set to a value slightly larger than 1 so that the activation has a constant value.

4.1. Structure of the Proposed Method

The novelty of OI-NMF lies in the modeling of the binary masks and combination with the onsets. In this section, the modeling of key valiables and an inference algorithm of OI-NMF are explained.

4.1.1. Binary Mask

The prior distribution of the binary mask $\boldsymbol{S}$ is modeled using a Markov chain on the assumption that a musical sound lasts a particular duration depending on the type of musical instrument. When the musical sound is present and its activation has a certain value, the value of the binary mask is 1 (on state). Accordingly, when it is absent and its activation has a value close to 0, the value is 0 (off state). Each element of the binary mask transitions in on/off states depending on the state of the previous time frame. Here, $a_{1\to1}, a_{0\to1} \in (0,1)$ are respectively a transition probability of on to o state and off to on state, and an initial probability $a_0 \in (0,1)$ determines the state of the first time frame. The assumption of sound-sustain characteristics is that the same state of a Markov chain is likely to continue. Therefore, $a_{1\to1}$ and $1-a_{0\to1}$ are set close to 1. The first time frames of the binary mask can be in either state, a_0 is set to $1/2$.

In this model, the joint probability of each component of the binary mask $\boldsymbol{S}_k (k = 1, 2, \ldots, K)$ is represented as follows:

$$p(\boldsymbol{S}_k) = p(S_{k,1}) \prod_{t=2}^{T} p\left(S_{k,t} \mid S_{k,t-1}\right). \tag{8}$$

Then the joint probability of the whole binary mask $\boldsymbol{S}$ is derived by the Equation (8):

$$p(\boldsymbol{S}) = \prod_{k=1}^{K} p(\boldsymbol{S}_k) = \prod_{k=1}^{K} p(S_{k,1}) \prod_{t=2}^{T} p\left(S_{k,t} \mid S_{k,t-1}\right), \tag{9}$$

where $p(S_{k,1})$ is a probability distribution that each element of the first time frame of each component of the binary mask follows. Since each element of the binary mask must be 0 or 1, $p(S_{k,1})$ is formulated by a Bernoulli distribution:

$$p(S_{k,1}) = \text{Bernoulli}\left(S_{k,1} \mid a_0\right). \tag{10}$$

Similarly, $p\left(S_{k,t} \mid S_{k,t-1}\right)$ is a probability distribution that each element of the binary mask ($t \geq 2$) follows, and is represented as a product of two Bernoulli distributions:

$$p\left(S_{k,t} \mid S_{k,t-1}\right) = \text{Bernoulli}\left(S_{k,t} \mid a_{1\to1}\right)^{S_{k,t-1}} \cdot \text{Bernoulli}\left(S_{k,t} \mid a_{0\to1}\right)^{1-S_{k,t-1}} \tag{11}$$

4.1.2. Onset

Assuming that the target instrument has $J (J < K)$ pitches, for each pitch $j = 1, 2, \ldots, J$, the corresponding onset time sequence $\boldsymbol{\tau}_j = (\tau_{j,1}, \ldots, \tau_{j,n}, \ldots, \tau_{j,N_j})$ is given, where N_j is the number of onset for pitch j. Each $\boldsymbol{\tau}_j$ is first given in time units and converted to time frames in the time-frequency domain. For simplicity, the onset sequence is formulated in the form of a matrix $\boldsymbol{O} \in \{0, 1\}^{K \times T}$, which is the same size of the binary mask;

$$O_{k,t} = \begin{cases} 1, & \tau_{k,n} \leq t \leq \tau_{k,n} + T_{\text{onset}} \\ & (k = 1, 2, \ldots, J, n = 1, 2, \ldots, N_j), \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

where T_{onset} is the tolerance width of the onset. If there is no tolerance (i.e. $T_{\text{onset}} = 1$), extraction will fail when the onset is given before the sound of the target instrument. To solve this problem, a certain width of tolerance is effective for robust separation. If T_{onset} is too long, the non-target sound behind the given onset will be overlapped. Therefore, we empirically adopted 1/8 beat length, which is the lower bound for the inference to work well, while trying ... 1/16, 1/8, 1/4, ... This onset matrix is utilized in the model inference described bellow.

4.2. Inference of the Proposed Model

In this model, in order to obtain the output variables (dictionary $\boldsymbol{W}$, activation $\boldsymbol{H}$, and binary mask $\boldsymbol{S}$), these posterior distributions should be calculated following Bayes rule. However, it is difficult to solve analytically because of these normalization terms. Therefore, posteriors are approximated by the expected value calculated by Gibbs sampling in reference to the BP-NMF Gibbs sampling algorithm [27]. In Gibbs sampling framework, each random variable is sampled from a conditional distribution given other variables. Each i-th random variable is sampled from the following conditional distributions:

$$\boldsymbol{W}^{(i)} \sim p\left(\boldsymbol{W} \,\middle|\, \boldsymbol{H}^{(i)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}\right), \quad (13)$$

$$\boldsymbol{H}^{(i)} \sim p\left(\boldsymbol{H} \,\middle|\, \boldsymbol{W}^{(i+1)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}\right), \quad (14)$$

$$\boldsymbol{S}^{(i)} \sim p\left(\boldsymbol{S} \,\middle|\, \boldsymbol{W}^{(i+1)}, \boldsymbol{H}^{(i+1)}, \boldsymbol{X}\right). \quad (15)$$

4.2.1. Sampling of the Binary Mask

Considering that the value of each element of the binary mask $S_{k,t}$ is binary, $S_{k,t}$ can be sampled following a Bernoulli distribution

$$S_{k,t} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X} \sim \text{Bernoulli}\left(S_{k,t} \,\middle|\, \frac{P_1}{P_1 + P_0}\right), \quad (16)$$

where likelihoods P_1, P_0 are represented as follows using $S_{\neg k,t}$ (all elements of $\boldsymbol{S}$ except $S_{k,t}$)

$$P_1 = p\left(S_{k,t} = 1 \mid S_{\neg k,t}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X}\right), \quad (17)$$

$$P_0 = p\left(S_{k,t} = 0 \mid S_{\neg k,t}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X}\right). \quad (18)$$

Equation (17) is divided into two terms

$$P_1 \propto p(S_{k,t} = 1) p\left(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{k,t} = 1, S_{\neg k,t}\right). \quad (19)$$

The first term of the Equation (19) is

$$p(S_{k,t} = 1) = \begin{cases} a_0, & t = 1 \\ a_{1\to1}^{S_{k,t-1}} a_{0\to1}^{1-S_{k,t-1}}, & t \geq 2 \end{cases}, \quad (20)$$

Algorithm 1 Gibbs Sampling for the Proposed Method

1: Initialize $\boldsymbol{W}, \boldsymbol{H}$ and $\boldsymbol{S}$
2: **for** $i = 1, 2, \cdots$ **do**
3: Calculate $\phi_{f,t,k} = \frac{W_{f,k} H_{k,t} S_{k,t}}{\sum_l W_{f,l} H_{l,t} S_{l,t}}$
4: Sample $\boldsymbol{W}$ following Equation (25)
5: Sample $\boldsymbol{H}$ following Equation (26)
6: Sample $\boldsymbol{S}$ following Equations (16), (23) and (24)
7: **end for**
8: Return expectation of $\boldsymbol{W}, \boldsymbol{H}$ and $\boldsymbol{S}$

and the second term is

$$p_{k,t}^1 \triangleq p\left(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{k,t} = 1, S_{\neg k,t}\right) \quad (21)$$

$$\propto \prod_{f=1}^{F} (X_{f,t}^{\neg k} + W_{f,k} H_{k,t})^{X_{f,t}} \exp(-W_{f,k} H_{k,t}), \quad (22)$$

where $X_{f,t}^{\neg k} = \sum_{l \neq k} W_{f,l} H_{l,t} S_{l,t}$. Thus

$$P_1 = \begin{cases} a_0 p_{k,t}^1, & t = 1 \\ a_{1\to1}^{S_{k,t-1}} a_{0\to1}^{1-S_{k,t-1}} p_{k,t}^1, & t \geq 2 \end{cases}. \quad (23)$$

Similarly, P_0 can be derived as follows:

$$P_0 = \begin{cases} (1 - a_0) p_{k,t}^0, & t = 1 \\ (1 - a_{1\to1}^{S_{k,t-1}})(1 - a_{0\to1}^{1-S_{k,t-1}}) p_{k,t}^0, & t \geq 2 \end{cases}, \quad (24)$$

where $p_{k,t}^0 \triangleq \prod_{f=1}^{F} (X_{f,t}^{\neg k})^{X_{f,t}}$. Note that the element of $\boldsymbol{S}$ to which the onsets are given, that is, the index where $O_{k,t} = 1$, is not sampled and fixed as 1. This is because the binary mask of the frame given the onsets can be considered to be in the on state. For others indexed, $S_{k,t}$ can be sampled from $t = 1$ following Equations (23), (24) and (16) in order.

4.2.2. Sampling of the Other Variables

Conditionals of the other variables, the dictionary $\boldsymbol{W}$ and the activation $\boldsymbol{H}$, are derived just like Gibbs sampling of BP-NMF [27] based on the conjugation,

$$W_{f,k} \mid \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{X} \sim \text{Gamma}\left(\alpha^W + \sum_{t=1}^{T} X_{f,t} \phi_{f,t,k}, \; \beta^W + \sum_{t=1}^{T} H_{k,t} S_{k,t}\right), \quad (25)$$

$$H_{k,t} \mid \boldsymbol{W}, \boldsymbol{S}, \boldsymbol{X} \sim \text{Gamma}\left(\alpha^H + \sum_{f=1}^{F} X_{f,t} \phi_{f,t,k}, \; \beta^H + S_{k,t} \sum_{f=1}^{F} W_{f,k}\right). \quad (26)$$

4.2.3. Sampling Algorithm for the Proposed Method

Algorithm 1 shows a Gibbs sampling algorithm of the proposed model. First, $\boldsymbol{W}, \boldsymbol{H}$, and $\boldsymbol{S}$ are initialized. It is known that the probability distribution inferred by Gibbs sampling converges to a stationary distribution regardless of the initial values. However, which spectral pattern appears on which component depends

largely on the initial values. Therefore, in order to induce the target instrument source to appear on the components given the onset, $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{S}$ are initialized according to the following rules referring to score-informed NMF [16, 17]. The dictionary $\boldsymbol{W}$ is randomly initialized following its prior distribution (Equation (5)). Assuming that musical sounds exist for an element given onsets, the components of the activation $\boldsymbol{H}$ given the onset are initialized to the expectation of the gamma distribution and 0. That is,

$$H_{k,t} = \begin{cases} \frac{\alpha^H}{\beta^H}, & O_{k,t} = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{27}$$

All other components (not given the onset) are initialized according to the gamma distribution. Similarly, the components of the binary mask $\boldsymbol{S}$ given the onsets are initialized as follows:

$$S_{k,t} = \begin{cases} 1, & O_{k,t} = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{28}$$

All the components not given the onsets are initialized to 1 because musical sound may exist in any time frame. After that, samples are drawn iteratively following its conditional distribution. The output values of $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{S}$ are obtained by taking the empirical average of the sample sequence after burn-in as their expectation. Note that burn-in is the period during which samples are rejected because it has not been reached the stationary distribution.

4.3. Reconstruction of the Target Source

Using the output value of $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{S}$ obtained by Gibbs sampling, the target instrument source is reconstructed in a manner similar to the standard NMF explained in section 3. As described in Section 4.2.1, the onsets imposed on the binary mask induces the corresponding sound to be sampled on the component given these onsets. Therefore, $\boldsymbol{W}_{\text{target}}$ and $\boldsymbol{H}_{\text{target}}$ can be constructed by using the all components given the onsets, that is, $k = 1, 2, \ldots, J$:

$$\boldsymbol{W}_{\text{target}} = \boldsymbol{W}_{:,1:J}, \tag{29}$$

$$\boldsymbol{H}_{\text{target}} = \boldsymbol{H}_{1:J,:} \circ \boldsymbol{S}_{1:J,:}. \tag{30}$$

Then, as in Equation 2, the spectrogram of the target instrument is recovered by Wiener filter, and the target signal is reconstructed by inverse STFT.

5. EXPERIMENTAL EVALUATION

To confirm that the proposed method can correctly separate a target instrument sound from a polyphonic sound using some of its onsets, we conducted an experiment to separate the melody line of the target instrument and evaluated its performance.

5.1. Simulation Settings

The dataset used in the experiment comprised musical pieces from MedleyDB, a realistic recording sound source dataset for sound source separation evaluation [28]. We selected eight jazz pieces from music produced by MusicDelta (BebopJazz, CoolJazz, FreeJazz, FunkJazz, FusionJazz, LatinJazz, ModalJazz, SwingJazz), that do not include vocals and have melody annotations of a dominant instrument. A 20 s excerpt from the head of the wav file of these pieces is taken and resampled at 22,050 Hz. Then, magnitude spectrograms were obtained by performing STFT of Hanning window and 512 FFT samples with 50% overlap. The harmonic part of the magnitude spectrogram obtained by harmonic/percussive source separation (HPSS) [29] is input to OI-NMF. The sample size of HPSS median filter is 31.

The input onsets used in the experiment were artificially generated from the F0 annotation of the target instrument included in the dataset. This annotation consisted of F0 value and timestamp. F0 value is converted to a MIDI note number, and recorded as time frames where the note number was switched as onsets. Note that F0 switching frames such as vibrato were excluded. Recorded onsets were input as described in Section 4.

Model inference was performed under the following conditions. The components number of OI-NMF was set to $K = 25$ as a sufficiently large value. The hyperparameters of the prior distributions were empirically set as follows: $\alpha^W = 0.5, \beta^W = 1.0, \alpha^H = 1.1, \beta^H = 1.0, a_0 = 0.5, a_{1\to1} = 0.99, a_{0\to1} = 0.01$. We executed 200 iterations Gibbs sampling with 100 burn-in to obtain 100 samples of each OI-NMF output variables. The expected values of these samples were output.

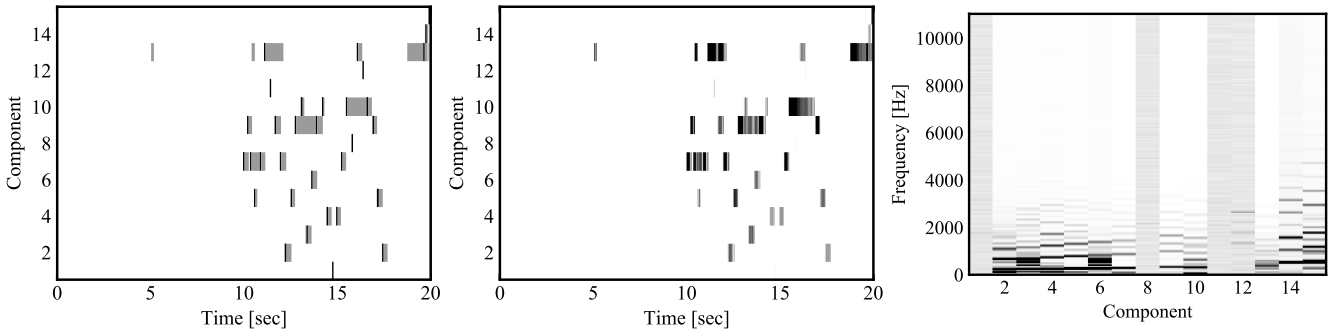
To evaluate the separation performance, we adopted scale-invariant metrics, signal-to-distortion ratio (SI-SDR), signal-to-interference ratio (SI-SIR), and signal-to-artifacts ratio (SI-SAR) [30]. SI-SDR is defined by the ratio between the estimated signal and the overall residual noise, and its larger value indicates higher separation performance. The residual noise is divided into the interference noise derived from the non-target sources and the artifacts noise, from the algorithm. SI-SIR and SI-SAR are defined by the ratio between the target and these noises. Since there is a trade-off between SI-SIR and SI-SAR, it is possible to grasp which noise is dominating in the separated sound.

Generally, scale-variant versions (SDR, SIR, and SAR) [31] are often used to evaluate the accuracy of blind source separation. However, when the estimated target source includes many silent sections as in the case of OI-NMF, these metrics cannot evaluate the separation accuracy correctly due to the scale change of the separated sound. Therefore, we think that scale-invariant metrics, which are improved versions to overcome the scale problem, are more appropriate for the evaluation of OI-NMF.

5.2. Separation Example

First, we show a separation example of OI-NMF. This example is from the SwingJazz piece in which a clarinet plays the melody. The whole onsets were used in this experiment. Figure 2 shows heatmaps of the OI-NMF output variables in this example. Figure 2 (a) shows the input onsets and the inferred binary mask. The binary mask was inferred like a piano roll from onsets in each component. Components 1, 8, 11 and 12 are those in which non-clarinet sound was estimated for which, the binary mask was 0 in frames other than onsets. Figure 2 (b) shows the element-wise product of the binary mask and activation. In the non-clarinet component, activation values were smaller than other components. Figure 2 (c) shows the dictionary. Since the input onsets are given in order from the lowest pitch, it was confirmed that components with corresponding harmonic structure appeared in order. Although non-clarinet components had a noisy structure inappropriate for instrument sound, those corresponding activations had small values that were almost negated. The improvement of SI-SDR in this example was about 4 dB. Other examples and each separated sound are published in the following repository[1].

[1] https://github.com/YutaKusaka/onset-informed-NMF-example



(a) Clarinet's onsets and binary mask $\boldsymbol{S}$ (b) Activation and binary mask $\boldsymbol{H} \circ \boldsymbol{S}$ (c) Dictionary $\boldsymbol{W}$

Figure 2: NMF output variables of the separation example (SwingJazz). (a) onsets (black lines) and heatmap of the NMF binary mask (gray bars). (b) heatmap of the element-wise product of the NMF activation and binary mask. (c) heatmap of the NMF dictionary.

5.3. Identification Experiment for Onset Effectiveness

To show that OI-NMF achieves the target instrument source separation by virtue of using onsets, we compare its separation performance with other methods without side-information. For comparison, two types of NMF model without side-information are used: OI-NMF without onsets input and Bayesian NMF [32], standard stochastic NMF model inferred by Gibbs sampling. As described in Section 1, these methods have no way of detecting which components correspond to the target instrument (i.e. all instruments appear on random components). Therefore, the recovered signal using the inferred component $k = 1, 2, \ldots, J$ in these methods gives the lower bound of the NMF separation performance and we can show the effectiveness of the onsets as side-information. The same melody separation experiment as in Section 5.2 is performed for each piece and method. Note that the experimental parameters of the methods without side-information are the same.

Table 1 shows a summary of the comparison. 10 separation trials were performed for each method and piece, and the mean and standard deviation of SI-SDR improvement from the mixture were analyzed for each piece and methods. We can confirm that SI-SDR is improved for all pieces in OI-NMF with onsets, that is, the melody separations were performed well. On the other hand, the means are around 0 or negative and the standard deviations are large for all pieces in the two methods without onsets. This is because the target instrument appears on random components and the separation is not done as expected.

Table 1: SI-SDR improvement [dB] between the method with and without the onsets. EachS value outside the parentheses indicates the mean and inside, the standard deviation.

	OI-NMF		Bayesian NMF
	with onsets	without onsets	
Bebop	**5.62** (2.46)	−2.75 (3.32)	−4.42 (5.80)
Cool	**4.84** (2.75)	−0.56 (3.10)	−0.83 (5.38)
Free	**4.49** (1.63)	−3.37 (5.24)	−8.44 (14.7)
Funk	**9.86** (0.97)	−3.69 (3.99)	−4.84 (7.32)
Fusion	**7.09** (1.21)	−1.54 (3.33)	0.33 (3.22)
Latin	**5.77** (0.37)	−4.58 (11.9)	−6.67 (10.3)
Modal	**4.52** (1.92)	−5.60 (4.05)	−1.77 (5.52)
Swing	**4.29** (1.09)	−2.38 (2.36)	−6.08 (1.82)

5.4. Verification Experiment for Onsets Absence

When users create onsets, it is expected that some onsets are lacking because they missed listening. To verify the effect of the absent input onsets, the melody separation was conducted with the proportion 100%, 75%, 50%, and 25% of the onset presence. The result of 100% is the same as in Section 5.3, and for each proportion other than 100%, 10 trials were also conducted and metrics were analyzed.

Figure 3 shows a summary of the experiments. Figure 3 confirms that the target sound could be separated in all pieces under all conditions because mean SI-SDR improvement was positive. In Figure 3, the mean tends to decrease, and the standard deviation increase with increasing the onset absence. There were some trials where the target instrument was not sampled correctly in the components given the onsets probably due to absence. In this case, the target instrument cannot be separated well. For this reason, we presume these tendencies occur. Regarding SI-SIR and SI-SAR, they indicate a similar tendency as SI-SDR. Comparing these metrics, we can confirm that the noise from the algorithm is dominant over the interference.

To examine the performance of the OI-NMF under the worst onset conditions, an experiment in which only one onset was present in each component was conducted. In this experiment, SI-SDR improvement was not achieved in many pieces. In particular, separation failed 4 / 10 times in FreeJazz, 6 / 10 in LatinJazz and ModalJazz, and 8 / 10 in SwingJazz. Considering these results, if too few onsets are input, OI-NMF often fails to perform separation. On the other hand, from the previous experiments, we consider that OI-NMF can achieve robust separation and tolerates the lack of a few onsets.

5.5. Verification Experiment for Onsets Error

In addition to the onset absence, it is expected that the onsets are deviated from the true onsets position due to human perception. We conducted an additional experiment to investigate the effect of this deviation on the separation performance. The deviation is modeled based on the studies on the generation of the onsets. The onset's deviation is assumed to follow a normal distribution, with an average of 10 ms [33] and a standard deviation of 100 ms [34]. The deviated onsets $\tilde{\tau}$ is expressed as follows with respect to the original onsets τ,

$$\tilde{\tau} = \tau + \epsilon, \tag{31}$$

$$\epsilon \sim \mathcal{N}(0.01, 0.1^2). \tag{32}$$

We compared the separation performance when ground truth 100% onsets and 100% onsets with above deviation. Other experimental settings are the same as the previous experiments.

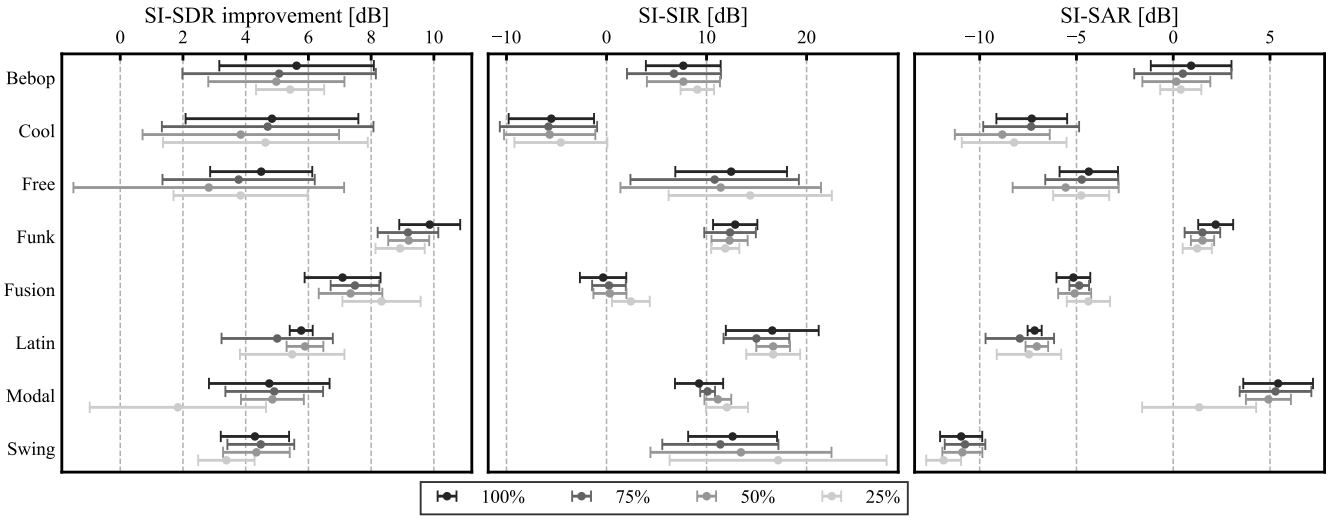


Figure 3: Separation results of each musical piece with the proportion of onsets presence. The dots show mean and error bars represent standard deviations. The legend shows the proportion of the onset presence.

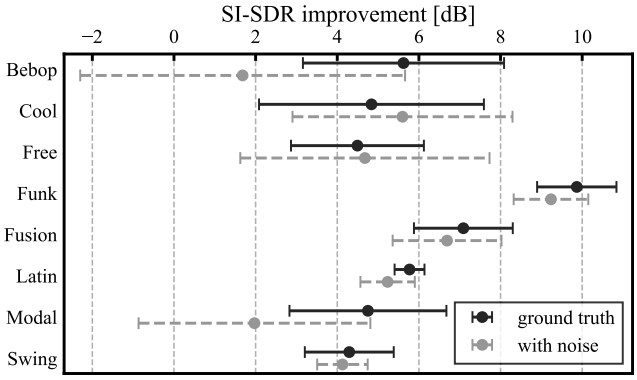


Figure 4: SI-SDR improvement when the input onsets are deviated. "Ground truth" and "with noise" show the results with and without the onset deviation respectively. The dots show the mean and the error bars represent the standard deviations.

Figure 4 shows the experiment results. From Figure 4, it is confirmed that for all pieces except CoolJazz and FreeJazz, the separation performance decreases due to the deviation of onsets. The improvement of SI-SDR of BebopJazz and ModalJazz are much lower than that of its ground truth, while the degradation of the performance is suppressed in other pieces. The major cause of the performance degradation in BebopJazz and ModalJazz is that the noisy components of the accompaniment instruments are likely to be estimated instead of the target due to the deviation. This noisy component spreads over music, causing a significant deterioration in performance (Figure 5). In CoolJazz and FreeJazz, the separation performance may be improved compared to the ground truth. It can be considered that the backward-deviated onsets are more appropriate than the ones we created from F0 annotations for these pieces.

In summary, these experimental results indicate the effectiveness of the proposed model and onsets for the target instrument separation. In addition, OI-NMF is expected to perform the robust separation against the absence and deviation of the onsets at the level expected when created by humans.

6. CONCLUSION AND DISCUSSION

In this paper, we proposed a new sound source separation framework, named onset-informed NMF, that can separate a target instrument sound from polyphonic sound by using target instrument onsets as side-information. To make use of onsets in the NMF source separation framework, the conventional NMF model is expanded by overlaying a binary mask based on a Markov-model on an NMF activation and treated onsets in its structure. Furthermore, we improved the inference algorithm to infer NMF variables including the binary mask and the onsets. Experiments to verify and assess its performance in target sound source separatoin with missing data showed OI-NMF could separate a target instrument sound without requiring the whole of its onsets.

As a current problem, in situations when the target and other sources onset simultaneously (e.g. a bass activates together with a piano melody), OI-NMF may incorrectly estimate these sources into one component due to the nature of NMF. Therefore, better separation can be expected by including constraints based on the target source structure such that the pitch difference in the same instrument is expressed by a shift of its harmonic structure in the frequency direction. Moreover, the wrong onset input or the type of the target instrument probably influence the separation performance. We plan to carry out experiments for these problems.

In addition, input onsets are currently created from the annotation in the dataset and assumed to be given by grouping for the same pitch. If users create onsets based on this assumption, the operation to group onsets can be time-consuming. Therefore, we are planning to construct an improved framework that can utilize onsets as a time series without pitch information.

7. REFERENCES

[1] Kazuyoshi Yoshii et al., "Drumix: An audio player with real-time drum-part rearrangement functions for active music listening," *Information and Media Technologies*, 2 (2), pp. 601–611, 2007.

[2] A. J. Simpson et al., "Deep karaoke: Extracting vocals

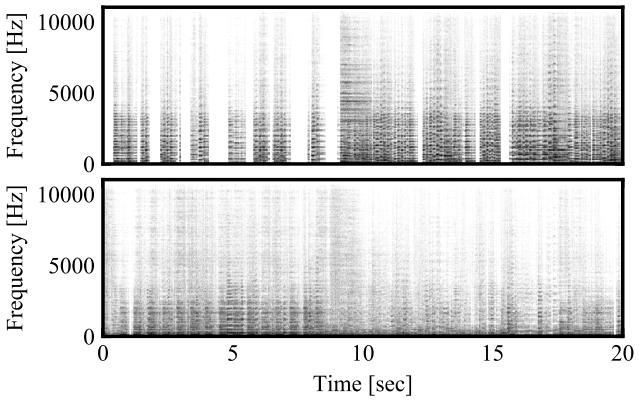


Figure 5: Reconstructed target (upper) and accompaniment (lower) spectrograms of BebopJazz with onset deviation.

from musical mixtures using a convolutional deep neural network," *LVA/ICA*, 2015, pp. 429–436.

[3] E. Benetos et al., "Automatic music transcription: challenges and future directions," *J. Intell. Inf. Syst.*, 41 (3), pp. 407–434, 2013.

[4] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," *ICASSP*, 2000, vol. 2, pp. II753–II756.

[5] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proc. IEEE*, 92 (4), pp. 712–729, 2004.

[6] M. Marolt, "A mid-level representation for melody-based retrieval in audio collections," *IEEE Trans. Multimedia.*, 10 (8), pp. 1617–1625, 2008.

[7] S. S. Shwartz et al., "Robust temporal and spectral modeling for query by melody," *SIGIR*, 2002, pp. 331–338.

[8] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," *NIPS*, 2001, pp. 556–562.

[9] C. Févotte et al., "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Comput.*, 21 (3), pp. 793–830, 2009.

[10] A. Hyvräinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, 13 (4), pp. 411–430, 2000.

[11] A. Liutkus et al., "An overview of informed audio source separation," *WIAMIS*, 2013, pp. 1–4.

[12] P. Smaragdis and G. J. Mysore, "Separation by "humming": User-guided sound extraction from monophonic mixtures," *WASPAA*, pp. 69–72.

[13] A. Lefèvre et al., "Semi-supervised NMF with time-frequency annotations for single-channel source separation," *ISMIR*, 2012, pp. 1–6.

[14] D. Kitamura et al., "Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing," *ISSPIT*, 2013, pp. 392–397.

[15] D. Kitamura et al., "Music signal separation by supervised nonnegative matrix factorization with basis deformation," *DSP*, 2013, pp. 1–6.

[16] S. Ewert and M. Muller, "Using score-informed constraints for NMF-based source separation," *ICASSP*, 2012, pp. 129–132.

[17] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," *ICASSP*, 2013, pp. 888–891.

[18] S. Uhlich et al., "Deep neural network based instrument extraction from music," *ICASSP*, 2015, pp. 2135–2139.

[19] P. Chandna et al., "Monoaural audio source separation using deep convolutional neural networks," *LVA/ICA*, 2017, vol. 10169, pp. 258–266.

[20] A. Ozerov et al., "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," *ICASSP*, 2011, pp. 257–260.

[21] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, 401 (6755), pp. 788–791, 1999.

[22] B. Wang and M. D. Plumbley, "Musical audio stream separation by non-negative matrix factorization," *DMRN*, 2005, pp. 1–5.

[23] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio Speech Lang. Process*, 15 (3), pp. 1066–1074, 2007.

[24] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," *WASPAA*, 2003, vol. 3, pp. 177–180.

[25] D. W. Griffin and J. S. Lim, "Signal Estimation From Modified Short-Time Fourier Transform.," *ICASSP*, 1983, vol. 2, pp. 804–807.

[26] D. Liang et al., "Beta Process Sparse Nonnegative Matrix Factorization for Music," *ISMIR*, 2013, pp. 375–380.

[27] D. Liang and M. D. Hoffman, "Beta Process Non-negative Matrix Factorization with Stochastic Structured Mean-Field Variational Inference," *arXiv:1411.1804*, 2014, pp. 1–6.

[28] R. Bittner et al., "MedleyDB: A multitrack dataset for annotation-intensive MIR research," *ISMIR*, 2014, pp. 155–160.

[29] D. FitzGerald, "Harmonic/Percussive Separation Using Median Filtering," *DAFx*, 2010, pp. 1–4.

[30] J. L. Roux et al., "SDR - half-baked or well done?," *ICASSP*, 2019, pp. 626–630.

[31] E. Vincent et al., "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio Speech Lang. Process*, 14 (4), pp. 1462–1469, 2006.

[32] A. T. Cemgil, "Bayesian Inference for Nonnegative Matrix Factorisation Models," *Comput. Intell. Neurosci.*, 2009, pp. 1–17, 2009.

[33] P. Leveau et al., "Methodology and tools for the evaluation of automatic onset detection algorithms in music.," *ISMIR*, 2004, pp. 1–4.

[34] M. Cartwright et al., "Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations," *Proc. ACM Hum. Comput. Interact.*, 1, pp. 1–21, 2017.