Nonnegative Matrix Partial Co-Factorization for Spectral and Temporal Drum Source Separation

Minje Kim, Jiho Yoo, Kyeongok Kang, and Seungjin Choi, *Member, IEEE*

***Abstract*—We address a problem of separating drum sources from monaural mixtures of polyphonic music containing various pitched instruments as well as drums. We consider a spectrogram of music, described by a matrix where each row is associated with intensities of a frequency over time. We employ a joint decomposition to several spectrogram matrices that include two or more column-blocks of the mixture spectrograms (columns of mixture spectrograms are partitioned into 2 or more blocks) and a drum-only (drum solo playing) matrix constructed from various drums *a priori*. To this end, we apply nonnegative matrix partial co-factorization (NMPCF) to these target matrices, in which column-blocks of mixture spectrograms and the drum-only matrix are jointly decomposed, sharing a factor matrix partially, in order to determine common basis vectors that capture the spectral and temporal characteristics of drum sources. Common basis vectors learned by NMPCF capture spectral patterns of drums since they are shared in the decomposition of the drum-only matrix and accommodate temporal patterns of drums because repetitive characteristics are captured by factorizing column-blocks of mixture spectrograms (each of which is associated with different time periods). Experimental results on real-world commercial music signal demonstrate the performance of the proposed method.**

***Index Terms*—Blind source separation, music source separation, nonnegative matrix factorization, nonnegative matrix partial co-factorization.**

I. Introduction

MUSIC source separation (MSS) attracts many related applications these days. The principal applications of MSS can be found in making quality Karaoke accompaniments and discovering higher-level information lying in music signals.

A. Karaoke Applications of MSS

First of all, as the Karaoke market is concerned with higher quality recorded sound instead of traditional MIDI based accompaniments, needs for adequate vocal and drum source separation are growing. Nowadays, many popular songs are recomposed and recorded again by the Karaoke service providers which are more possibly selected by the end-users. It is obvious that the end-users prefer re-recorded accompaniment sounds to the traditional MIDI sounds, because of the more realistic sound quality. However, it is also true that the service providers suffer higher cost of making those realistic accompaniment sounds than that of making MIDI based ones. Consequently, MSS can be an alternative solution to securing quality accompaniment sounds when it meets the decent separation performance that the market requires. One of the main goals of MSS in this sense is to take away main vocal sources from the commercial music without harming the quality of the remaining sound; it should contain all the other instrumental sources and chorus while the main vocal source should be repressed. On top of that, the accompaniment sound should not contain too much artifacts, which usually caused by deliberate source separation processes.

Another important goal of MSS for the Karaoke application is to separate drum sources from the mixtures. Drum sources do affect the user experiences of the Karaoke service as well as vocal sources, since they play a great role in maintaining the quality of the sound when the users want to change certain features of the accompaniment sounds, such as key. For example, separated drum sources can remain the same while the other harmonic instruments change their key in order to fit the users' voice range; thus the attack of each drum can convey the same rhythmic atmosphere (atmosphere that is usually expressed by repeating attacks or beats of percussive instruments) of the song without changing its frequency characteristics. Similarly, object-based audio services [1] and their standard [2] also require clearly segregated music sources, since the main goal of the services is to let the end-users to experience their own configuration of instruments by allowing them to control certain features of each instrument, such as volume.

B. MSS in Music Information Retrieval

We can find other interesting usages of MSS in the music information retrieval (MIR) field. There have been efforts to enhance the accuracy of audio classification tasks by using rhythm and bass-line features. The basic intuition about these is that patterns of rhythm sources and bass guitars are important factors that are responsible for recognizing musical genres and moods that music triggers. Simple rhythm and bass-line feature extraction methods were involved into audio genre [3] and music mood [4] classification tasks. Those

This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2010, NIPA ITRC Support Program (NIPA-2011-C1090-1131-0009), and NRF World Class University Program (R31-10100).

Minje Kim and Kyeongok Kang are with the Realistic Acoustics Research Team, Electronics and Telecommunications Research Institute (ETRI), 138 Gajeongno, Yuseong-gu, Daejeon, 305-700, Korea (email: {mkim, kokang}@etri.re.kr).

Jiho Yoo is with the Department of Computer Science, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang, Kyungbuk 790-784, Korea (email: zentasis@postech.ac.kr).

Seungjin Choi is with the Department of Computer Science and the Division of IT Convergence Engineering, Pohang University of Science and Technology, San 31, Hyoja-dong, Nam-gu, Pohang, Kyungbuk 790-784, Korea (email: seungjin@postech.ac.kr).

methods were not intended to isolate rhythm or bass guitar signals from the mixture, but showed prospective usage of DSS in audio classification tasks. One reason why the referred works could not adapt sophisticated DSS methods can be that audio classification tasks usually process a bunch of audio signals for their training and test procedure, where complex feature extraction steps, including source separation, might delay whole classification procedure consequently. However, it is quite promising that more isolated music sources can help extract more meaningful features. Another usage of DSS in MIR is to enhance tempo detection [5], where probabilistic latent component analysis (PLCA) [6] was used to separate rhythmic sources. The approach showed that proper DSS can help improve a certain MIR task, tempo detection. Similarly, the drum transcription performance was also improved by adequate drum isolation [7].

C. Characteristics of Rhythmic Sources

Most of modern commercial songs usually contain drum tracks to convey rhythmic atmosphere of the songs. Aside from acoustic drum sets that are being used widely, various types of instruments, even with pitches, can also be grouped into rhythmic sources, such as claps, triangles, and timpani. Therefore, it is true that there is no definite criterion whether an instrument is a rhythm instrument or not. However, it is also clear that traditional drum sets, which consist of kick drum, snare, tom-toms, hi-hat, and several cymbals, are representative rhythmic instruments. With some exceptions above, in this clause, we would like to survey some dominant characteristics of drum sources.

- Spectral characteristics: less harmonious (more noisy) and more energy in high frequency bands.
- Temporal characteristics: more impulsive, faster decaying time, more periodic, and more repeating.

If we consider each note with individual frequency from a harmonic instrument as a distinctive source, its excitations are not repeating enough compared with the one for a drum instrument. It is obvious that a harmonic instrument can be played during a whole song; it can be seen that it is repeating at a glance. However, the problem lies in the fact that the particular instrument cannot be factorized well enough to allocate exactly one particular basis vector to one specific note. Although it could be possible, the note-related basis vector seldom appears.

To summarize, relatively small number of distinctive frequency characteristics, which appear often enough to be considered that they are repeating, are required of a rhythmic instrument. For convenience sake, we use the term drum source separation to refer rhythmic source separation as well.

D. Drum Source Separation by Matrix Factorization

Although DSS can be seen as a kind of challenging tasks similarly to the other single channel source separation problems, DSS can be easier than others since drum sounds have properties which other sound sources do not have to this extent, e.g. the ”spectral flatness (noiselikeness)” and ”repeatability,” which are incorporated in this paper.

Because spectral noiselikeness and temporal impulsiveness of drum instruments usually construct vertical ridges, spectrogram analysis can distinguish them from parallel ridges of pitched instruments. Maximum A Posteriori (MAP) based separation systems were developed by using those characteristics of each group of source [8] [9]. Also, distinctive time varying gains of harmonic and stochastic signals were employed to enhance subband-based decomposition method [7]. In that, improved Wiener filtering method based on more elaborated spectral shaping of drum sources using NMF and adaptation was proposed. In addition, gamma chain priors were employed in the domain of tensor factorization models to take difference of temporal continuity of pitched / unpitched sources into account [10]. Although the referred works are good examples of incorporating temporal properties of sources, an attack followed by fast decaying or lack of temporal continuity, other important temporal characteristics, namely repeatability, still need to be addressed.

By regarding the input magnitude spectrograms as matrices to decompose, the factor matrices of independent subspace analysis (ISA) were used to extract several rhythmic features, and then classified with heuristically developed rules to distinguish drums and harmonic sources [11]. A succeeding work [12] using nonnegative matrix factorization (NMF) instead of ISA was proposed where the resulting basis and corresponding encoding vectors of NMF were classified into two groups: drums and harmonic sources. After the basis vectors are classified using support vector machines (SVM), those in the drums group can be used to reconstruct drum sources. Although their approach showed quite acceptable separation results, the assumption that each decomposed basis vector should represent either only drum or harmony sources is not sufficient to separate real-world music signals. In other words, the standard NMF lacks separation performance about a specific source, which eventually causes the consequent degradation of basis classification accuracy.

To tackle this problem, nonnegative matrix partial co-factorization (NMPCF) was introduced to separate drum sources from the commercial music mixtures. There were two kinds of NMPCF-based approaches to DSS. The first one explicitly exploited prior knowledge of drum sources [13]. In [13], a solo playing of various drums was used as an auxiliary input signal along with the mixture signal to be separated. The drum sources consisting the solo playing were taken from irrelevant songs to the input mixture signals. The main premise of the referred work is that NMPCF jointly decomposes those two input signals by sharing some basis vectors of the two factorization tasks. After all, the resulting commonly shared basis vectors represent both the whole drum solo playing and drum sources in the mixture signal.

The second approach using NMPCF targets at separating rhythmic sources from the mixture without any explicit prior knowledge of drum sources [14]. Therefore, it involved a temporal property, namely repeatability, that the most drum sources have, while the other harmonic sources cannot be represented well by repeating excitations of reasonably small number of basis vectors. This is because of the fact that the drum sources usually do not change their frequency char-

acteristics while pitched instruments alter their notes. Thus, distinguishing repetitive sound components can be a way to separate drum sources. To this end, the system segmented the only input matrix, the magnitude spectrogram of the mixture signal, into column-blocks, and then found out common basis vectors among them using NMPCF. The task was called a *rhythmic* source separation, since not only drum sources, but some harmonic sources can also be separated when their frequency variations are not significant and are repeating enough to convey rhythmic atmosphere of the song, for example, bass guitars.

In this paper, we propose a unified approach to harmonize those two branches of NMPCF-based DSS systems. As the classification system in [11] first seeks to extract spectral and temporal features, respectively, drum and harmonic sources clearly differs in those two types of characteristics. On the other hand, each NMPCF-based DSS system only exploits one of them, spectral properties in [13] or temporal properties in [14].

E. Organization of the Paper

Section II of this paper reviews predecessors of DSS. Next, Section III provides the proposed unified system. In addition, experimental results in Section IV show that the consolidated system performs better than the previous works. Finally, Section V concludes this work.

II. Previous Works on Drum Source Separation

In this section, we address distinctive characteristics of drum sources that are used in preceding DSS systems. Afterward, a classification system using standard NMF is reviewed, which influenced the advent of its successors. On top of that, NMPCF-based DSS systems are also introduced with their drawbacks that the proposed system desires to attack.

A. NMF-Based Music Source Separation Systems

NMF was first introduced as a dimensionality reduction method which desired to infer sparse representations of an input nonnegative matrix [15]. As a matrix decomposition scheme, one of the advantages of NMF is that it successfully finds out nonnegative components of the input matrix, whose additive reconstruction manner allows sparser representation of the input matrix, like the same way as the human brain processes the information. NMF seeks to find out two factor matrices that reconstruct the input matrix by well-designed multiplicative update rules, which enforce the factor matrices to remain nonnegative during and after the learning process if they are initialized with nonnegative random values [16].

NMF [17] and its extensions with shift-invariance concept [18][19][20] have been utilized as a tool to factorize a magnitude spectrogram of a input mixture signal. They inspired many researchers by their preliminarily results where the monophonic input spectrograms were factorized well. On top of that, picking up some of the basis vectors as building blocks for reconstructing desired target sources, was a relatively simple way to separate music sources given only one mixture signal.

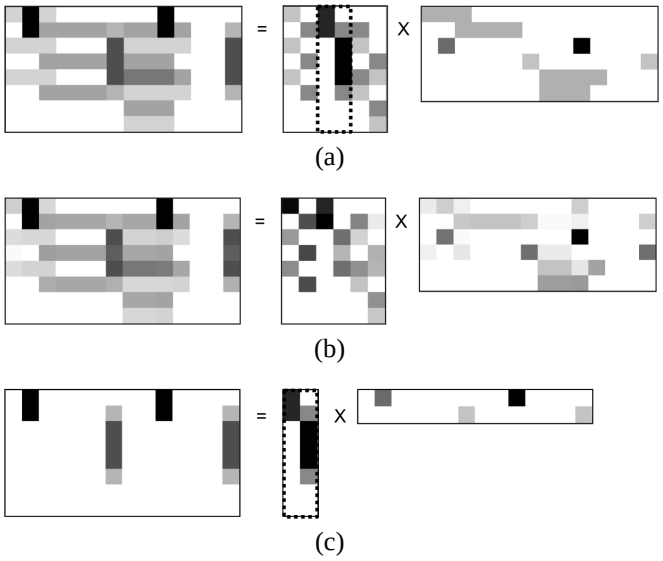


Fig. 1: Pictorial examples of magnitude spectrogram decomposition. (a) Ideal decomposition consists of six basis vectors. The two basis vectors of toy drum sources are marked with dotted box. (b) Empirical decomposition results with NMF. (c) Reconstructed toy drum sources by selecting proper basis vectors of NMF decomposition results. The selected basis vectors are marked with dotted box, which are same with the ones in (a).

NMF seeks to reduce the difference between the product of a couple of factor matrices and the input matrix:

$$\boldsymbol{X} \approx \boldsymbol{AS}.$$

If we regard the input matrix $\boldsymbol{X}^{M\times N} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]$ as the magnitude spectrogram of the mixture signal, the resulting nonnegative factor matrices $\boldsymbol{A}^{M\times R} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_R]$ and $\boldsymbol{S}^{R\times N} = [\boldsymbol{s}_1, \boldsymbol{s}_1, ...\boldsymbol{s}_R]^\top$ can represent the spectral basis vectors and their temporal encodings, respectively.

Fig. 1 gives us a pictorial example of decomposing frequency magnitude spectrogram. Fig. 1 (a) is a case of an ideal decomposition, where six basis (column) vectors and encoding (row) vectors represent both spectral characteristics and temporal excitations of the input matrix, respectively. Specifically, the first two basis vectors represent two notes of an harmonic instrument and the last two basis vectors are for two notes of another harmonic instrument with different timbre, while the third and forth ones represent a toy drum set.

In order to separate a particular source from the mixture in Fig. 1 (a), a selection mechanism is needed. If one can distinguish a subset of basis vectors or their encodings which compound the target source, just summating all the outer-products of the selected basis vectors and their corresponding intensity vectors is sufficient to reconstruct the target source. For example, the mixture matrix in Fig. 1 (a) consists of three sources. If we select first two basis vectors and their corresponding encodings, the magnitude spectrogram of the first source $\boldsymbol{Y}_1$ can be recovered like this:

$$\boldsymbol{Y}_1 \approx \boldsymbol{a}_1\boldsymbol{s}_1^\top + \boldsymbol{a}_2\boldsymbol{s}_2^\top.$$

The selection of two basis vectors for the source $\boldsymbol{Y}_1$ in this case is based on the intuition that the timbre, which is represented as a harmonics structure, is not changed even though the source can play different notes. The first and second basis vectors share same harmonics structure while they differ in their positions to reflect their different pitches. Likewise, the third source consists of the fifth and sixth basis vectors, whose harmonics structures are quite different from the ones of the first source. Furthermore, the third and forth basis vectors, which are marked with dotted box, do not contain any harmonics structures and appear abruptly; they epitomize the characteristics of drum sources. Selecting those basis vectors and corresponding encodings can lead the drum sources-specific decomposition in Fig. 1 (c). The referred systems in [18][19] [20] involved an extended version of NMF to capture the spectrally shifting common shape of a particular source.

On the other hand, Fig. 1 (b) gives us an exemplary failure that NMF can make. While the forth and sixth basis vectors in Fig. 1 (a) stand for a drum and a note of played by the third source, the corresponding ones in Fig. 1 (b) fail to estimate the original ones. This is because of two facts: there can be multiple solutions to decompose a matrix and the learning process of NMF can converge in the local minima.

B. NMF-based Drum Bases Classification

In Fig. 1, we noticed that proper selection of basis vectors is a key step to MSS in NMF-based frameworks. However, manual ordering of the basis vectors is almost impossible when there are dozens of basis vectors to be ordered. Classification systems on ISA [11] or NMF [12] bases were proposed to tackle this problem. Fig. 2 provides the block diagram of the referred NMF-bases classification [12]. The system consists of two tracks: training and separation procedures. As for the training procedure, two different types of 10 seconds-long training signals are collected and decomposed using standard NMF, separately. The authors set 20 and 10 for the number of basis vectors of the two classes, harmony and drum instruments. Since each excerpt in each class is learned individually, the resulting 20 or 10 basis vectors solely represent the class-specific characteristics. Afterward, feature extraction step is devoted to refine NMF resulting factor matrices with higher-level features, such as MFCC, spectral centroid, roll-off point, periodicity, and so on. Those features are then used as input of the SVM training step. In the separation procedure, NMF works a little differently since the mixture excerpts are fed into it instead of already labeled sources; it decomposes an input spectrogram with 30 basis vectors. After feature extraction step, the SVM classifier learned in the training procedure classifies the basis vectors according to their features. Basis vectors which classified to drum instrument class, are then used to synthesize drum sources that are mixed in the given input mixture signal.

The main cause of reconstruction error of the referred system [12] is that the standard NMF does not provide a *separation-friendly* decomposition. Like in the case of Fig. 1 (b), MSS with NMF can fail to perfectly separate even with

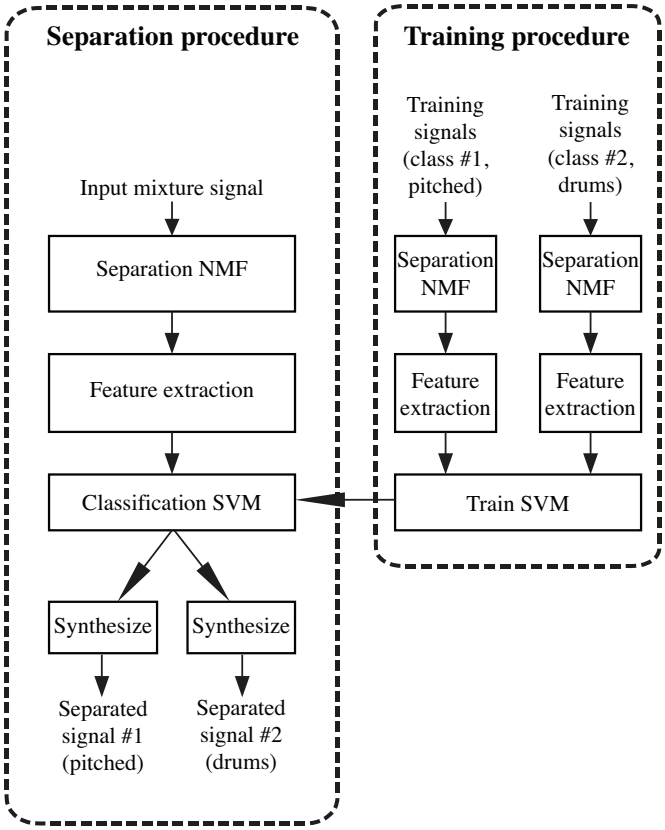


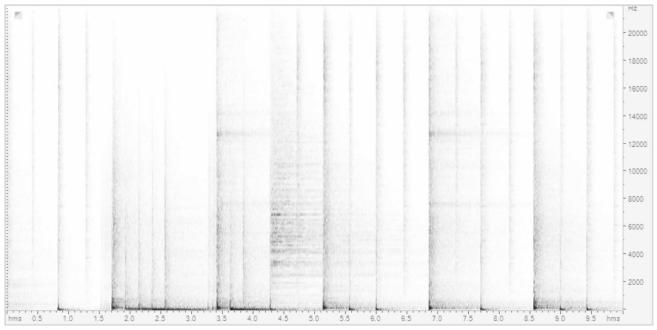
Fig. 2: The block diagram of the referred system [12].

simple configuration of instruments, for example, two to three sources not playing too many notes, simultaneously. If the number of sources and notes increases, it is almost impossible to learn basis vectors which contain pure frequency characteristics of only one source at a time. Therefore, although it is true that the referred system [12] does not strictly assume its individual basis vectors to represent only one source, some of the contaminated basis vectors are apt to be misclassfied. Fig. 3 support this argument clearer. Compared with the original drum source in Fig. 3 (a) and (b) shows an almost perfect reconstruction, since the classification in Fig. 3 (b) was done with deliberately permuted 20 basis vectors of harmonic instruments and 10 basis vectors of drum instruments: only one out of the ten drum basis vectors was misclassified into harmonic instrument class. That means that if factorization step perfectly purifies the basis vectors, reordering those permuted basis vectors using feature extraction and classification methods is not a big deal.

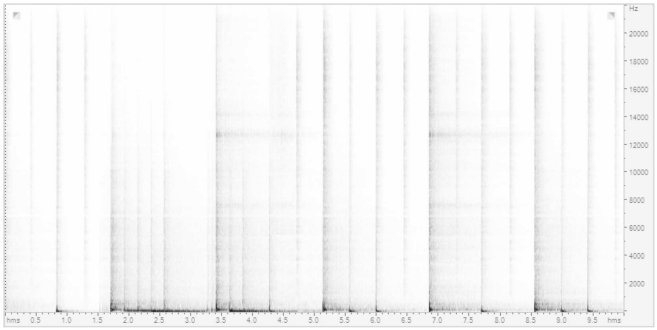
C. NMPCF-based Drum Source Separation

Nonnegative matrix partial co-factorization (NMPCF) emerged from the concept of joint decomposition or collective matrix factorization, which let the multiple input matrices be decomposed into several factor matrices while some of them are shared [21][22][23][24][25].

1) NMPCF for Spectral Drum Source Separation: As a DSS algorithm, NMPCF was introduced to co-factorize a mixture signal and a prior side information [13]. In other words,



(a)



(b)

Fig. 3: An ideal classification result with separately learned NMF basis vectors. (a) Original spectrogram of a drum source. (b) Reconstruction of the drum source using 9 basis vectors classified to drum sources except only one missclassified basis vector out of ten.

some basis vectors of a factor matrix of the mixture signal are also used to factorize the side information matrix. For given two input matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, the joint decomposition approximates them with following models:

$$\boldsymbol{X}^{(1)} \approx \boldsymbol{A}_C \boldsymbol{S}_C^{(1)} \quad (1)$$

$$\boldsymbol{X}^{(2)} \approx \boldsymbol{A}_C \boldsymbol{S}_C^{(2)} + \boldsymbol{A}_I^{(2)} \boldsymbol{S}_I^{(2)}. \quad (2)$$

Fig. 4 (a) describes the two models in (1) and (2). $\boldsymbol{X}^{(1)}$ is the side information matrix, which consists of solo playing of various drum sets as prior knowledge. Therefore, the model in (1) is dedicated to reconstruct the input matrix $\boldsymbol{X}^{(1)}$ by letting R_C number of the basis vectors, $\boldsymbol{A}_C^{M \times R_C}$, contain spectral shapes of the drum-only input magnitude spectrogram. On the other hand, the second factorization model in (2) for approximating mixture signal to be separated can be described with two parts: common and individual decompositions. Without considering the former model in (1), the distinction of two parts are meaningless. However, by commonly sharing the same basis vectors $\boldsymbol{A}_C^{M \times R_C}$, which also utilizes spectral information of drum-only signal, the first term in (2), $\boldsymbol{A}_C \boldsymbol{S}_C^{(2)}$, partly reconstructs drum-like sources that are mixed in the second input matrix $\boldsymbol{X}^{(2)}$, too. Contrarily, the individual basis vectors $\boldsymbol{A}_I^{(2)M \times R_I}$ are devoted to represent the other sources, which exist only in the second input matrix $\boldsymbol{X}^{(2)}$. To summarize, the first term $\boldsymbol{A}_C \boldsymbol{S}_C^{(2)}$ can partly reconstruct the sources, which shares similar spectral characteristics to the ones in $\boldsymbol{X}^{(1)}$, while the second term $\boldsymbol{A}_I \boldsymbol{S}_I^{(2)}$ is responsible for the other sources, which are spectrally not similar to the ones in $\boldsymbol{X}^{(1)}$. In DSS case, for example, the common characteristics of spectral bases would be high noiselikeness and more energy in high frequency bands. Therefore, we propose to call this method *NMPCF for Spectral DSS (S-DSS)*.

The benefits of NMPCF for S-DSS compared with NMF plus SVM scheme in clause II-B are twofold. First of all, it does not require complicated feature extraction and classifier learning steps. Instead, during and after NMPCF learning process basis vectors are ordered into two groups, each of them represents one of two sets of sources. Secondly, the resulting two sets of basis vectors are more separation-friendly; they are more likely to be distinctive than those of standard NMF. As we have seen in Fig. 1 (b) and Fig. 3, NMF is not responsible for distinguishing its resulting basis vectors into groups or refining them to fit into a desirable source; the only criterion for NMF to factorize is nonnegativity of the factor matrices. On the other hand, NMPCF for S-DSS additionally tries to make common basis vectors $\boldsymbol{A}_C$ to reflect the spectral property of the auxiliary input signal $\boldsymbol{X}^{(1)}$. After all, $\boldsymbol{A}_C$ can not only be grouped during the learning process, but be discriminative.

The main drawback of this scheme is that it does not incorporate with temporal characteristics of the target source, since the common basis vectors of NMPCF for S-DSS are designed to reflect spectral ones only. It is obvious that involving the unique temporal properties of drum sources, such as repeatability, will help enhance the separate performance.

2) NMPCF for Temporal Drum Source Separation: At the same time, another usage of NMPCF was introduced to separate repeating sources [14]. Instead of involving a prior knowledge input, the scheme tries to find out repeating sources, such as drums, by partially co-factorizaing column-blocks of mixture matrix based on the assumption that there will be common basis vectors that represent those repeating sound components across all the blocks. Furthermore, it also attempts to provide alternative solution to DSS when there is no decent prior knowledge signals about the drum sources. Since it only utilizes a temporal property of drum sources, repeatability, we propose to call this method *NMPCF for Temporal DSS (T-DSS)*. The number of factorization models of NMPCF for T-DSS varies with the number of column-blocks, but, with a simple two segments case, they can be expressed like,

$$\boldsymbol{X}^{(1)} \approx \boldsymbol{A}_C \boldsymbol{S}_C^{(1)} + \boldsymbol{A}_I^{(1)} \boldsymbol{S}_I^{(1)} \quad (3)$$

$$\boldsymbol{X}^{(2)} \approx \boldsymbol{A}_C \boldsymbol{S}_C^{(2)} + \boldsymbol{A}_I^{(2)} \boldsymbol{S}_I^{(2)}. \quad (4)$$

Fig. 4 (b) despicts this situation. For a given input mixture matrix $\boldsymbol{X}$, a proper number of segmentation is done to produce L consecutive column-blocks (in this case, $L = 2$). Since there is no auxiliary input matrix to be taken into consideration, the common basis vectors of all the blocks $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ will be learned to contain commonly appearing spectral components throughout the whole blocks, since the input column-blocks are exclusive time periods of the mixture signal. On the other hand, similarly to the one in (2), $\boldsymbol{A}_I^{(1)}$ and $\boldsymbol{A}_I^{(2)}$ are responsible for recovering the other harmonic instruments that lie in each

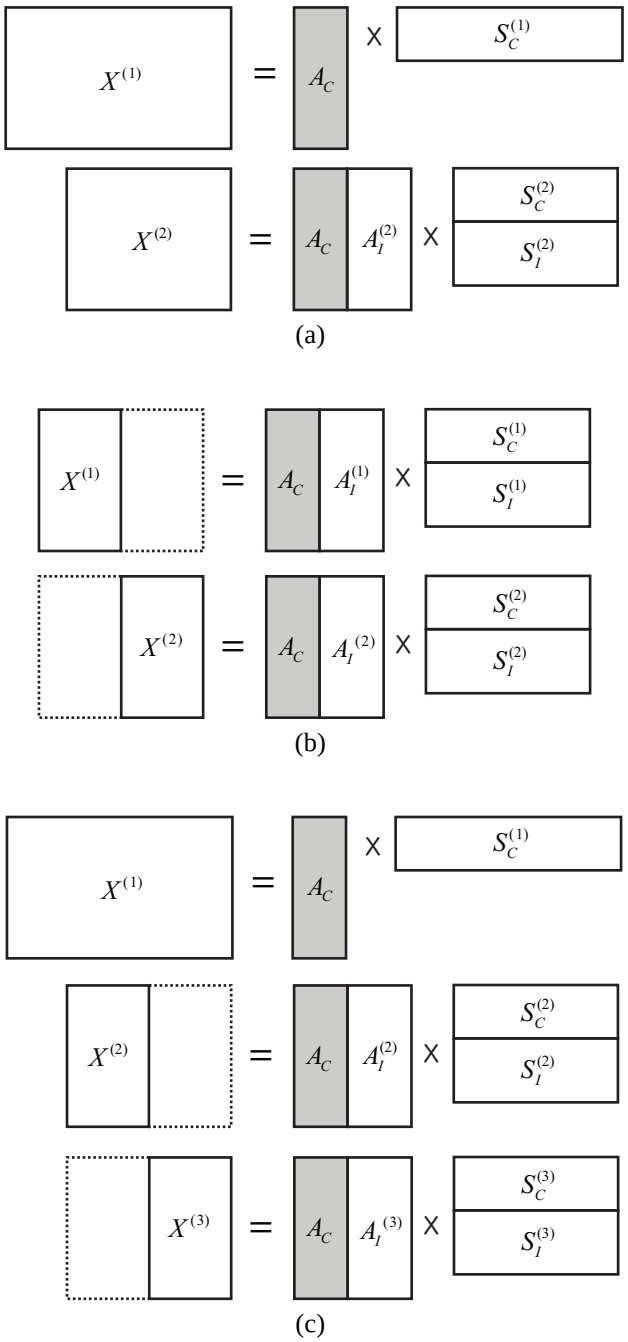


Fig. 4: Pictorial illustration of matrix decomposition models using NMPCF. (a) Spectral DSS model with solo playing of various drums and mixture signal as two input matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, respectively. (b) Temporal DSS model without any prior knowledge input matrix. The only input matrix, a magnitude spectrogram of the mixture signal, is segmented into multiple excerpts (column-blocks) and fed to NMPCF learning process. This figure stands for the case with two column-blocks. (c) The proposed spectral and temporal DSS model. It involves both the prior knowledge matrix and segmented mixture matrices to make advantage of both spectral and temporal characteristics of target sources.

column-block with different configurations and various notes.

Although the reported separation performance of NMPCF

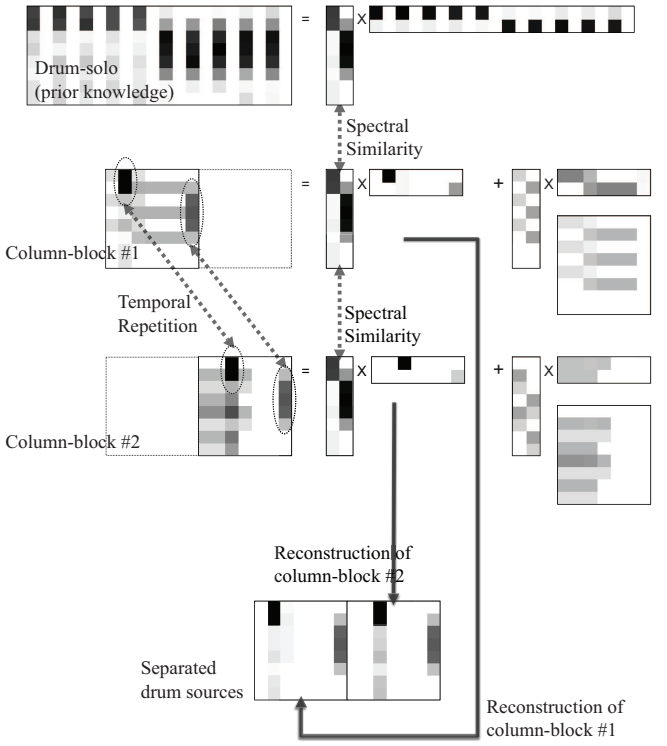


Fig. 5: Conceptual flow of the proposed DSS system. Noticeable spectral and temporal characteristics of drum sources are marked with dotted arrows.

for T-DSS is worse than S-DSS, it widens the applicability of NMPCF-based DSS systems to the case where no massive drum-solo playing is available. Also, it is promising that it might be able to recover some virtual rhythmic instruments whose dominant pitches prevent utilization of the supervised methods. Admittedly, the temporally repeating property is not enough to cover all the different spectrums of drum sources, either.

III. NMPCF for Spectral and Temporal Drum Source Separation

A. Overview of the Separation System

In this section, we propose a unified NMPCF solution to DSS system that can capture both the spectral and temporal characteristics of drum sources. This method can help reduce the flaws of the previous NMPCF-based DSS systems, S-DSS and T-DSS. Similarly, we named the proposed method *NMPCF for Spectral and Temporal DSS (ST-DSS)*.

Fig. 5 describes conceptual flow of ST-DSS system with a simple DSS task. We regard the input matrix of Fig. 1 (a) as the magnitude spectrogram of mixture signal to be separated as well. The goal of this separation task is to get toy drum spectrogram of Fig. 1 (c). To this end, an additional input matrix is utilized as a prior knowledge of our toy drums. It consists of several spectrally drum-like excitations. Therefore, we can factorize it and get two representative basis vectors. On the other hand, the mixture input matrix is partitioned into two column-blocks. If we simply factorize each block without any prior knowledge or interaction between them, the resulting

four basis vectors need to be classified to select proper drum-like basis vectors.

In the figure, however, we can easily notice that there is significant similarity between the basis vectors of drum-solo input matrix and two basis vectors of each column-block, marked with dotted vertical arrows. This similarity is the result of the effort of NMPCF learning process, which forces the basis vectors to reside in the designated positions, the first and second columns in the two decompositions of column-blocks and coerces them to be identical with the one for factorizing the drum-solo input. Eventually, we do not need to classify the basis vectors. Furthermore, they can also represent spectral shapes of drum sources in the mixture matrices, because they are intensively affected by the massive drum-solo input.

Additionally, there are repeating patterns across the two column-blocks, which are designated by dotted arrow and ellipses. During the repeating excitation, the spectral shapes are not changed. Therefore, sharing two basis vectors of the two block-wise decomposition can allow them to contain spectral shape of repeating sources.

Finally, the common basis vectors and their corresponding block-specific encodings reproduce the drum sources in block-wise manner. The partitioned reconstructions are then serialized to build the whole spectrogram. Compare the matrix of separated drum sources in Fig. 5 with the ideal one in Fig. 1 (c). Note that the other two basis vectors of each column-block are dedicated to reconstruct the other not-repeating notes of harmony sources.

The actual sharing of those specific basis vectors are implemented in the learning process by sharing same variables in the aggregated objective function and update rules, which are described in clause III-C.

B. Factorization Models

The factorization models of ST-DSS system let the common basis vectors $\boldsymbol{A}_C$ be shared by decompositions of prior knowledge matrix $\boldsymbol{X}^{(1)}$, and all the column-blocks of the input mixture matrix, like

$$\begin{aligned}\boldsymbol{X}^{(1)} &\approx \boldsymbol{A}_C\boldsymbol{S}_C^{(1)}\\ \boldsymbol{X}^{(2)} &\approx \boldsymbol{A}_C\boldsymbol{S}_C^{(2)}+\boldsymbol{A}_I^{(2)}\boldsymbol{S}_I^{(2)}\\ \boldsymbol{X}^{(3)} &\approx \boldsymbol{A}_C\boldsymbol{S}_C^{(3)}+\boldsymbol{A}_I^{(3)}\boldsymbol{S}_I^{(3)}.\end{aligned}$$

Fig. 4 (c) depicts those models with two column-blocks and a prior knowledge signal. Firstly, input matrix $\boldsymbol{X}^{(1)}$ stands for a prior side information about the drum sources. Specifically, it is the same one with $\boldsymbol{X}^{(1)}$ in Fig. 4 (a). Also, the column-blocks $\boldsymbol{X}^{(2)}$ and $\boldsymbol{X}^{(3)}$ along with their decompositions correspond to the ones in (3), (4) and in Fig. 4 (b). However, the main improvement of the new proposed models is that the common basis vectors $\boldsymbol{A}_C$ can represent both spectral characteristics of the input matrix $\boldsymbol{X}^{(1)}$ and temporally repeating components of all the column-blocks $\boldsymbol{X}^{(2)}$ and $\boldsymbol{X}^{(3)}$.

C. Objective Function and Update Rules

Objective function and update rules for previous NMPCF-based DSS systems were derived to cover general cases in [14]. In this clause, we further generalize the objective functions using the concept of β-divergence [26] [27].

For given L input nonnegative matrices $\boldsymbol{X}^{(l)}$ $(1 \le l \le L)$, NMPCF seeks to minimize

$$\begin{aligned}\mathcal{J}_{\text{NMPCF}} &= \textstyle\sum_{l=1}^{L}\lambda_l\left\|\boldsymbol{X}^{(l)}-\boldsymbol{A}_C\boldsymbol{S}_C^{(l)}-\boldsymbol{A}_I^{(l)}\boldsymbol{S}_I^{(l)}\right\|_F^2\\ &\quad+\gamma\left\{\textstyle\sum_{l=1}^{L}\left\|\boldsymbol{A}^{(l)}\right\|_F^2\right\}\\ &= \textstyle\sum_{l=1}^{L}\lambda_l\mathcal{D}_F\left(\boldsymbol{X}^{(l)}\middle|\,\boldsymbol{A}_C\boldsymbol{S}_C^{(l)}+\boldsymbol{A}_I^{(l)}\boldsymbol{S}_I^{(l)}\right)\\ &\quad+\gamma L\mathcal{D}_F\left(\boldsymbol{A}_C|\boldsymbol{0}\right)+\gamma\textstyle\sum_{l=1}^{L}\mathcal{D}_F\left(\boldsymbol{A}_I^{(l)}|\boldsymbol{0}\right),\end{aligned}$$

where the regularization term $\sum_{l=1}^{L}\left\|\boldsymbol{A}^{(l)}\right\|_F^2$ is defined by $\sum_{l=1}^{L}\left\|\boldsymbol{A}^{(l)}\right\|_F^2 = L\|\boldsymbol{A}_C\|_F^2+\sum_{l=1}^{L}\left\|\boldsymbol{A}_I^{(l)}\right\|_F^2$, and $\mathcal{D}_F(\boldsymbol{A}|\boldsymbol{B})$ represents the Frobenius norm of $(\boldsymbol{A}-\boldsymbol{B})$.

The β-divergence is defined as

$$\mathcal{D}_\beta(x|y)=\frac{1}{\beta(\beta-1)}\left(x^\beta+(\beta-1)y^\beta-\beta xy^{\beta-1}\right)$$

for $\beta \in \Re\setminus\{0,1\}$. For $\beta=0$ and $\beta=1$, the β-divergence can be defined as the limit value of the above definition, which becomes the Kullback-Leibler (KL) divergence

$$\mathcal{D}_{\beta=1}(x|y)=x(\log x-\log y)+(y-x),$$

for the case of $\beta=1$ and becomes the Itakura-Saito (IS) divergence

$$\mathcal{D}_{\beta=0}(x|y)=\frac{x}{y}-\log\frac{x}{y}-1,$$

for the case of $\beta=0$. The KL-divergence and IS-divergence were usually known to bring better result especially in the musical signal processing, possibly because of the scale invariance [28] of the divergence.

To apply the β-divergence on the objective function of NMPCF, we can obtain

$$\begin{aligned}\mathcal{J}^{\beta}_{\text{NMPCF}} &= \textstyle\sum_{l=1}^{L}\lambda_l\mathcal{D}_\beta\left(\boldsymbol{X}^{(l)}\middle|\,\boldsymbol{A}_C\boldsymbol{S}_C^{(l)}+\boldsymbol{A}_I^{(l)}\boldsymbol{S}_I^{(l)}\right)\\ &\quad+\gamma L\mathcal{D}_F\left(\boldsymbol{A}_C|\boldsymbol{0}\right)+\gamma\textstyle\sum_{l=1}^{L}\mathcal{D}_F\left(\boldsymbol{A}_I^{(l)}|\boldsymbol{0}\right).\end{aligned}\quad(5)$$

We can calculate the derivative of the $\mathcal{D}_\beta(x|y)$ with respect to y as

$$\frac{\partial\mathcal{D}_\beta(x|y)}{\partial y}=y^{\beta-2}(y-x),$$

and this leads the following gradients of the objective functions

$$\begin{aligned}
\frac{\partial \mathcal{J}^{\beta}_{\text{NMPCF}}}{\partial \boldsymbol{A}_C} &= \sum_{l=1}^{L} \lambda_l \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \right. \\
&\quad \left. \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} - \boldsymbol{X}^{(l)} \right) \right\} \left(\boldsymbol{S}_C^{(l)} \right)^{\top} \\
&\quad + 2\gamma L \boldsymbol{A}_C \\
\frac{\partial \mathcal{J}^{\beta}_{\text{NMPCF}}}{\partial \boldsymbol{A}_I^{(l)}} &= \lambda_l \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \right. \\
&\quad \left. \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} - \boldsymbol{X}^{(l)} \right) \right\} \left(\boldsymbol{S}_I^{(l)} \right)^{\top} \\
&\quad + 2\gamma \boldsymbol{A}_I^{(l)} \\
\frac{\partial \mathcal{J}^{\beta}_{\text{NMPCF}}}{\partial \boldsymbol{S}_C^{(l)}} &= \lambda_l \boldsymbol{A}_C^{\top} \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \right. \\
&\quad \left. \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} - \boldsymbol{X}^{(l)} \right) \right\} \\
\frac{\partial \mathcal{J}^{\beta}_{\text{NMPCF}}}{\partial \boldsymbol{S}_I^{(l)}} &= \lambda_l \left(\boldsymbol{A}_I^{(l)} \right)^{\top} \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \right. \\
&\quad \left. \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} - \boldsymbol{X}^{(l)} \right) \right\}.
\end{aligned}$$

Multiplicative update rules to learn $\boldsymbol{S}^{(l)} = [\boldsymbol{S}_C^{(l)}, \boldsymbol{S}_I^{(l)}]$, $\boldsymbol{A}_C$, and $\boldsymbol{A}_I^{(l)}$ can be derived by taking positive terms of the partial derivative of the objective function (5) as numerator of multiplication factor, while taking negative terms as denominator,

$$\begin{aligned}
\boldsymbol{A}_C &\leftarrow \boldsymbol{A}_C \odot \frac{\sum_{l=1}^{L} \lambda_l \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \boldsymbol{X}^{(l)} \right\} \left(\boldsymbol{S}_C^{(l)} \right)^{\top}}{\sum_{l=1}^{L} \lambda_l \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-1)} \left(\boldsymbol{S}_C^{(l)} \right)^{\top} + 2\gamma L \boldsymbol{A}_C} \\
\boldsymbol{A}_I^{(l)} &\leftarrow \boldsymbol{A}_I^{(l)} \odot \frac{\lambda_l \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \boldsymbol{X}^{(l)} \right\} \left(\boldsymbol{S}_I^{(l)} \right)^{\top}}{\lambda_l \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-1)} \left(\boldsymbol{S}_I^{(l)} \right)^{\top} + 2\gamma \boldsymbol{A}_I^{(l)}} \\
\boldsymbol{S}_C^{(l)} &\leftarrow \boldsymbol{S}_C^{(l)} \odot \frac{\boldsymbol{A}_C^{\top} \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \boldsymbol{X}^{(l)} \right\}}{\boldsymbol{A}_C^{\top} \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-1)}} \\
\boldsymbol{S}_I^{(l)} &\leftarrow \boldsymbol{S}_I^{(l)} \odot \frac{\left(\boldsymbol{A}_I^{(l)} \right)^{\top} \left\{ \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-2)} \odot \boldsymbol{X}^{(l)} \right\}}{\left(\boldsymbol{A}_I^{(l)} \right)^{\top} \left(\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)} \right)^{\cdot(\beta-1)}},
\end{aligned}$$

which can be simplified by

$$\begin{aligned}
\boldsymbol{A}_C &\leftarrow \boldsymbol{A}_C \odot \frac{\sum_{l=1}^{L} \lambda_l \left\{ \left(\boldsymbol{A}^{(l)} \boldsymbol{S}^{(l)} \right)^{\cdot(\beta-2)} \odot \boldsymbol{X}^{(l)} \right\} \left(\boldsymbol{S}_C^{(l)} \right)^{\top}}{\sum_{l=1}^{L} \lambda_l \left(\boldsymbol{A}^{(l)} \boldsymbol{S}^{(l)} \right)^{\cdot(\beta-1)} \left(\boldsymbol{S}_C^{(l)} \right)^{\top} + 2\gamma L \boldsymbol{A}_C} \\
\boldsymbol{A}_I^{(l)} &\leftarrow \boldsymbol{A}_I^{(l)} \odot \frac{\lambda_l \left\{ \left(\boldsymbol{A}^{(l)} \boldsymbol{S}^{(l)} \right)^{\cdot(\beta-2)} \odot \boldsymbol{X}^{(l)} \right\} \left(\boldsymbol{S}_I^{(l)} \right)^{\top}}{\lambda_l \left(\boldsymbol{A}^{(l)} \boldsymbol{S}^{(l)} \right)^{\cdot(\beta-1)} \left(\boldsymbol{S}_I^{(l)} \right)^{\top} + 2\gamma \boldsymbol{A}_I^{(l)}} \\
\boldsymbol{S}^{(l)} &\leftarrow \boldsymbol{S}^{(l)} \odot \frac{\left(\boldsymbol{A}^{(l)} \right)^{\top} \left\{ \left(\boldsymbol{A}^{(l)} \boldsymbol{S}^{(l)} \right)^{\cdot(\beta-2)} \odot \boldsymbol{X}^{(l)} \right\}}{\left(\boldsymbol{A}^{(l)} \right)^{\top} \left(\boldsymbol{A}^{(l)} \boldsymbol{S}^{(l)} \right)^{\cdot(\beta-1)}}.
\end{aligned} \tag{6}$$

Note that these update rules reduce to the Frobenius norm case when $\beta = 2$.

D. Separation Procedure

TABLE I describes the procedure of ST-DSS using NMPCF. Note that in S-DSS case, where no segmentation is made, there will be only two input matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, which stand for drum-solo prior knowledge matrix and the whole mixture spectrogram, respectively. On the other hand, if there is no prior knowledge matrix to be used, $\boldsymbol{X}^{(l)}$s are all consist of consecutive column-blocks of the input mixture spectrogram in T-DSS case. Finally, in ST-DSS case of TABLE I, where we decide to take advantage of both spectral and temporal properties of drum sources, the first input matrix should be magnitude spectrogram of a drum-solo signal, while the other $L-1$ matrices are the column-blocks of the mixture signals. In the case of S-DSS, and ST-DSS, $\boldsymbol{A}_I^{(1)}$ and $\boldsymbol{S}_I^{(1)}$ are all initialized with empty matrices.

TABLE I: Separation Procedure of ST-DSS

1) Prepare the input matrices
   a) Prepare the magnitude spectrogram of the drum solo-playings $\boldsymbol{X}^{(1)}$
   b) Prepare the magnitude spectrogram of the mixture signal, then segment it into predefined $L-1$ number of column-blocks, $\boldsymbol{X}^{(l)}$ $(2 \le l \le L)$
   c) Prepare phase information of all mixture column-blocks, $\Phi^{(l)}$ $(2 \le l \le L)$
   d) Initialize factor matrices $\boldsymbol{A}_C$, $\boldsymbol{A}_I^{(l)}$, and $\boldsymbol{S}_I^{(l)}$ $(2 \le l \le L)$ with random positive values
   e) Initialize factor matrices $\boldsymbol{S}_C^{(l)}$ $(1 \le l \le L)$ with random positive values
   f) Initialize factor matrices $\boldsymbol{A}_I^{(1)}$ and $\boldsymbol{S}_I^{(1)}$ with empty matrices
2) Update each factor matrix using (6) for the predefined number of iterations
3) Reconstruct the separated signals
   a) Reconstruct the spectrogram of drum sources in each column-block:
   $\boldsymbol{Y}^{(l)} = (\boldsymbol{A}_C \boldsymbol{S}_C^{(l)}) \odot \Phi^{(l)}$ $(2 \le l \le L)$
   Or alternatively, Wiener filtering can be used like:
   $\boldsymbol{Y}^{(l)} = \boldsymbol{X}^{(l)} \odot \frac{\boldsymbol{A}_C \boldsymbol{S}_C^{(l)}}{\boldsymbol{A}_C \boldsymbol{S}_C^{(l)} + \boldsymbol{A}_I^{(l)} \boldsymbol{S}_I^{(l)}}$ $(2 \le l \le L)$
   b) Concatenate column-block reconstructions:
   $\boldsymbol{Y} = [\boldsymbol{Y}^{(2)}, \boldsymbol{Y}^{(3)}, \cdots, \boldsymbol{Y}^{(L)}]$
   c) Inverse-transform the connected spectrogram into time-domain

Although there have been efforts to estimate phase values of each NMF component, reconstruction using only one basis vector and its corresponding encoding row vector [29] [30], it is still hard to learn them in NMPCF update process. Therefore, we set aside phase values of each colum-blocks of mixture spectrogram $\Phi^{(l)}$ $(2 \le l \le L)$, and simply reuse them to reconstruct drum sources of each block. Or alternatively, we can get more natural-sounding reconstructions by using a type of Wiener filtering like in [10].

IV. Experimental Results

A. Experimental Envionments

TABLE II summarizes the experimental environments.

- About songs: for the experiments, commercially released Korean pop songs are used to evaluate the DSS systems. Their genres are varied, including acoustic ballads with electric guitars, R&B, and electronic dance pop at various tempos. Also, the drum track of each song is secured to measure the quality of estimated one.
- Sampling rate and bit per sample: All the input signals are sampled at a rate of 44.1 kHz, and encoded in 16

TABLE II: Experimental environments.

	S-DSS	T-DSS	ST-DSS
Sampling rate / bit per sample	44.1kHz / 16 bit / Mono		
Transform	STFT (2048 or 4096), Hamming Window(2048 or 4096), Overlap(1792 or 3840)		
Number of test signals	10		
Length of each test signal	10 sec.		
Length of each segment	10 sec.	[0.5, 1, 2, 5] sec.	
Number of training signal	13	-	13
Length of each training signal	10 sec.	-	10 sec.
γ	1		
λ_l	λ_1=[0.001, 0.01, 0.1, 1, 5, 10], λ_l=1 (for $l > 1$)		
Number of basis vectors	[5, 10, 20, 50, 100, 200]		
Number of iterations	Collected every separation result until it reaches 100		

bits just as in the ordinary PCM encoding of commercial CDs, and then downmixed into the single channel.

- Transform: 2048 or 4096 points of input waveforms are windowed with Hamming window, and then transformed into time-frequency domain by short time Fourier transform (STFT). Succeeding frames are overlapped with the previous ones by 1792 (2048 - 256) or 3840 (4096 - 256) points. The small fixed hopsize, 256, is for higher temporal resolution to capture the attacks of the drum sources.
- Test songs: 10 different songs were used as test materials. 10 seconds-long excerpt of each song is carefully chosen to contain as many sources as possible, such as drums, vocals, and a variety of the other pitched instruments. For T-DSS and ST-DSS, each test signal is split into column-blocks again with various length, such as 0.5, 1, 2, and 5 second. Note that NMPCF for S-DSS does not require partitioning input signal into column-blocks.
- Training signal: 10 seconds-long drum tracks of 13 different songs, which eventually amount to 130 seconds-long signal, are collected as prior knowledge. No one of the 13 songs for training signal is same with the ones for test signals. Note that T-DSS does not use any training signal.
- γ: regularization parameter γ prevents basis vectors from convergence to too small values. However, since it does not seem to be important in comparison of the systems, we set it to one for all cases.
- λ_l: each λ_l controls the contribution of the reconstruction error of each input matrix among column-blocks and drum-solo playing. Although heuristic choices of λ_ls are proposed in [13] and [14], we tested two exemplar songs with various values of λ_l to judge the sensitivity of the NMPCF algorithms to various configurations of parameter sets (see clause IV-B for more detail).
- Number of iterations: one of main ambiguity of NMF and its branches as source separation tools is that the convergence of objective function does not guarantee best separation performance. Often, too much iterations result in overfitting, while not enough iterations do not provide good results. We basically iterate all NMPCF algorithms 100 times to investigate performance oscillations that can be caused by iterations.
- Number of basis vectors: another important ambiguity of NMF-related methods is to choose optimal number of basis vectors. Although it can be assumed that each basis vector represents a particular note of a specific instrument (or an attack of a specific drum in DSS cases), both estimating the number of notes in a song and representing them separately with those bases are difficult. Therefore, experiments in IV-B explore the effect of the various numbers of basis vectors, too.

We evaluated the separation performances of the NMPCF-based DSS systems similarly to *BSS_EVAL* toolbox [31]. For some of our important applications, such as high quality Karaoke and object-based audio services, the proposed three allowed distortions in [31], time-invariant gains, time-invariant filters, and time-varying filters, are not appropriate, since all the original sources are mixed without any scale factor or filter effects and the reconstructed source should be used as is, not having any permutation or scale ambiguity. Furthermore, with those applications, we equally care about reducing both interfering other sources and algorithmic artifacts. Therefore, we used following decomposition model,

$$\hat{s}(t) = s(t) + e_{errors}(t), \tag{7}$$

where $\hat{s}(t)$, $s(t)$, and $e_{errors}(t)$ are the estimated source, original source, and all the possible errors including interference and artifacts, respectively. Therefore, the source-to-distortion ratio can be defined like,

$$\text{SDR} := 10\log 10 \frac{\sum_t s(t)^2}{\sum_t e_{errors}(t)^2}. \tag{8}$$

B. Sensitivity to Parameter Configuration

In this clause, we evaluate the separation performance of each NMPCF-based algorithm by changing their configuration of parameters. The goal of these experiments is to compare the robustness of separation performances considering the general usage of the algorithms when users are not aware of the parameter configuration that is optimal for the particular input mixture. We tested each NMPCF algorithms by changing their configuration of parameters, especially for the case of Frobenius norm objective function (β = 2).

Fig. 6 compares separation performances of three NMPCF-based separation systems for the input, test song 1. Note that these experiments were done with 2048 point FFT. We can check the change of SDR values with specific configurations of parameters, namely the number of basis vectors. For instance,

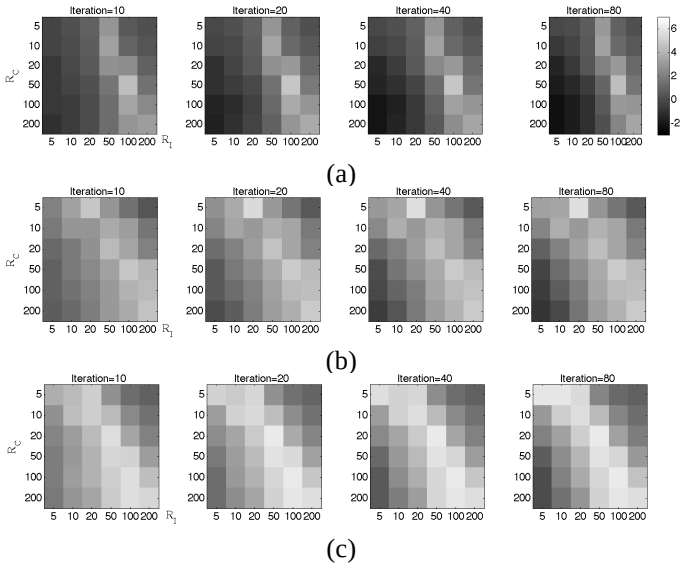


Fig. 6: SDR of the recovered drum source of test song 1 with 2048 point FFT. In each figure, x-axis is the number of individual basis vectors for harmonic sources and y-axis is the number of common basis vectors for drum (or rhythmic) sources, respectively. λ_1 and the length of column-block are fixed with the optimal ones. Brighter pixels represents higher SDR values. (a) SDR of S-DSS. λ_1 was fixed with its optimum for this song, 0.01. (b) SDR of T-DSS. The length of column-block was fixed with its optimum for this song, 2 sec. (c) SDR of ST-DSS. λ_1 and the length of column-block were fixed with their optimal values, 0.1 and 2 sec.

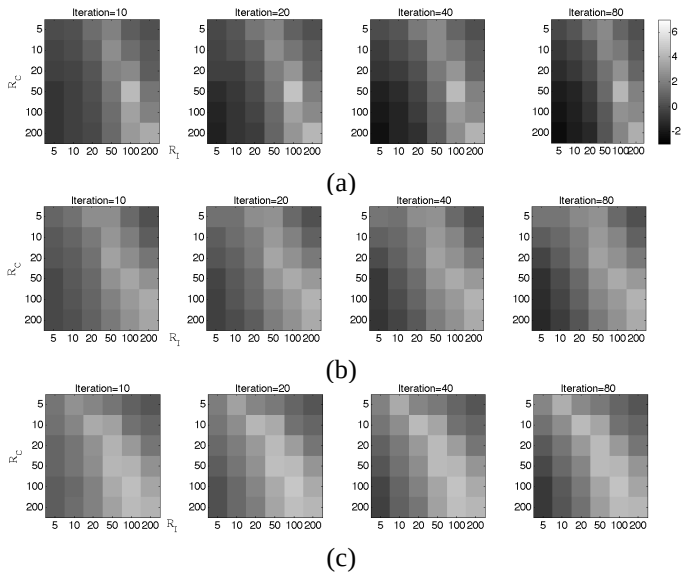


Fig. 7: SDR of the recovered drum source of test song 4 with 2048 point FFT. (a) SDR of S-DSS. $\lambda = 0.01$. (b) SDR of T-DSS. The length of column-block was fixed with 2 sec. (c) SDR of ST-DSS. $\lambda = 0.1$ and the length of column-block equals 2 sec.

we can see that T-DSS in Fig. 6 (b) has a wider bright region than S-DSS in (a), although in the particular case of configuration, $R_C = 50$, $R_I = 100$ and iteration 20,

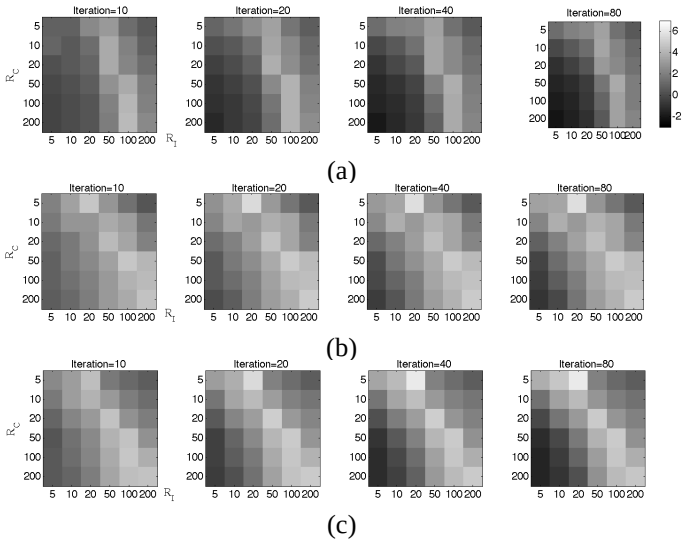


Fig. 8: SDR of the recovered drum source of test song 1 with 4096 point FFT. (a) SDR of S-DSS. $\lambda = 0.01$. (b) SDR of T-DSS. The length of column-block was fixed with 2 sec. (c) SDR of ST-DSS. $\lambda = 1$ and the length of column-block equals 2 sec.

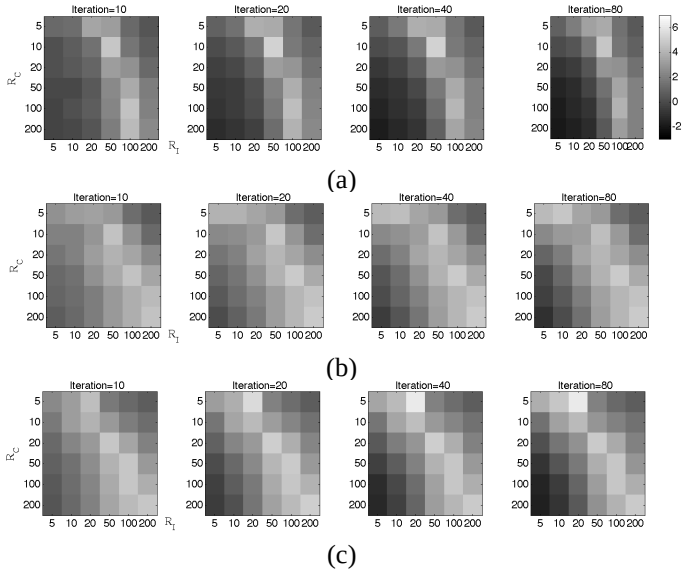


Fig. 9: SDRs of the recovered drum source of test song 4 with 4096 point FFT. (a) SDR of S-DSS. $\lambda = 0.1$. (b) SDR of T-DSS. The length of column-block was fixed with 2 sec. (c) SDR of ST-DSS. $\lambda = 1$ and the length of column-block equals 2 sec.

S-DSS performs better than all results of T-DSS. However, slightly different choice of configuration can cause serious degradation in the S-DSS case, because its bright region is so limited. Furthermore, ST-DSS raises the performance of T-DSS, while retaining its advantage of wide region of good configuration. In (c), we can check that the bright region is still large enough to cope with diverse choices of parameters and much brighter than that of T-DSS in (b). Therefore, if the selected configuration of numbers of basis vectors is not the optimal one, ST-DSS provides better results than S-DSS.

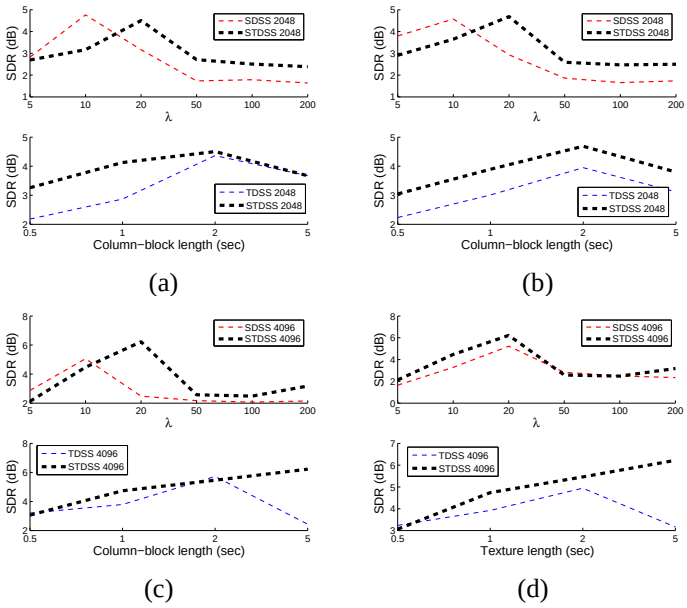


Fig. 10: Comparison of SDR results with varying parameters, λ_1 and the length of column block. Upper graphs represent results of S-DSS (thin red dashed lines) and ST-DSS (Thick black dashed lines) with changing λ_1. Bottom graphs represent results of T-DSS (thin blue dashed lines) and ST-DSS (Thick black dashed lines) with changing the length of column-block. (a) Results of song 1 with FFT size 2048 (b) Results of song 4 with FFT size 2048 (c) Results of song 1 with FFT size 4096 (c) Results of song 4 with FFT size 4096

Furthermore, ST-DSS results are generally superior than T-DSS, because they successfully incorporate with the advantage of using prior-knowledge about drum instruments, likely what S-DSS does.

Similarly, we can check the advantage of ST-DSS in the other experiments with different test song, number 4, in Fig. 7, and different FFT size, 4096, in Fig. 8 and Fig. 9.

By fixing R_C and R_I with appropriate optimal values, we can compare the effect of the other parameters, λ_1 and the length of column-block. The thick (black) dashed line in upper graph of Fig. 10 (a) represents SDR results of ST-DSS with varying λ_1. We can see that it demonstrates the wider good-performing region than S-DSS case, represented by thin (red) dashed line. Similarly, the other experiments with different input mixture, song 4, in (b) and with different FFT size, 4096, in (c) and (d), show that ST-DSS is less sensitive than or at least equally to S-DSS when the optimal λ_1 is not known. Additionally, we can also compare the influence of different length of column-block to ST-DSS and T-DSS systems. The bottom graphs of Fig. 10 (a), (b), (c), and (d) clearly show higher SDR values of ST-DSS than T-DSS (thin blue dashed line) with all experimental settings and inputs.

C. Separation Performances

Fig. 11 depicts the average SDR of 10 test songs recorded at each iteration. We evaluated them with three different NMPCF-based algorithms, ST-DSS, S-DSS, and T-DSS. Furthermore, for each DSS algorithm, three different objective functions were tested by changing β. It is not possible to assert that ST-DSS is always the best among the three, because the experiments were done with particular parameter configurations that are only the best for input test song 1. However, we can predict usual behaviors of each algorithm based on very general situation when users do not know a specific parameter configuration that is optimal for all kind of their test songs. Aside from the parameters discussed in the previous clause, another important decision point for users is the number of iteration. Although those three NMPCF-based algorithms are designed to minimize their own objective functions, minimization of objective functions does not always guarantee better separation results. Each different song has its own best number of iterations along with the optimal configuration of parameters. Therefore, if an algorithm stays in high SDR values during a wide range of iterations, the algorithm can be seen more useful in the practical applications.

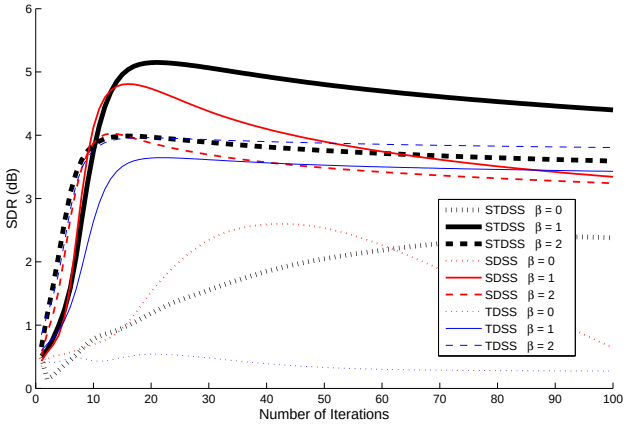
In Fig. 11 (a) and (b), where FFT sizes were set to 2048 and 4096, respectively, We can check that ST-DSS lines (the thickest black ones) formulate wider range of high SDR regions. For instance, users who fortunately stop those algorithms at 15th iteration in (a), they will get similar results from ST-DSS and S-DSS with $\beta = 1$. However, when they stop at arbitrary numbers of iterations, such as 40 and 60th, they cannot easily get good results from S-DSS while ST-DSS consistently provides good results than others. Similar characteristics of fast decaying of S-DSS can be seen in (b), too, with FFT size 4096.

We tried to compare the effect of β on drum source separation, but it was not clear with those experiments. That was because each test song has different optimal configuration of parameters when β differs.

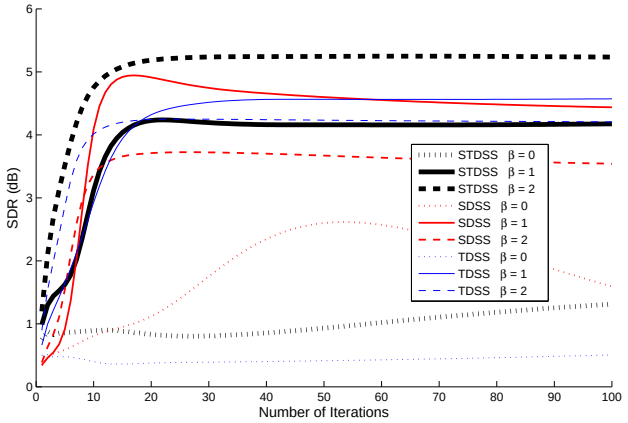
We can learn from Fig. 11 that users will get the best average SDR value around 20th iteration for this particular set of 10 test songs. However, a good separation algorithm should also deal with the other kind of user choice, namely

TABLE III: Separation performances of S-DSS and ST-DSS for the 10 commercial songs. Two objective functions based on two different values of $\beta = 1$ or 2 are assessed with a song-specific configuration of parameters. SDR results at iteration 20 and 60 are selected as representatives. FFT size was set to 2048.

	iteration=20		iteration=60	
	S-DSS	ST-DSS	S-DSS	ST-DSS
Song	β=1 / β=2	β=1 / β=2	β=1 / β=2	β=1 / β=2
1	2.35 / 4.68	4.10 / 3.16	1.34 / 4.51	4.00 / 3.10
2	3.91 / 3.67	4.45 / 2.68	1.63 / 1.90	3.62 / 2.20
3	6.36 / 4.76	5.89 / 4.86	5.50 / 4.46	5.54 / 4.45
4	3.19 / 3.95	4.01 / 3.21	1.92 / 3.55	3.98 / 3.16
5	4.17 / 1.99	5.86 / 4.82	4.61 / 2.20	5.30 / 4.53
6	6.06 / 4.28	6.56 / 6.75	5.41 / 4.36	5.94 / 6.57
7	3.86 / 3.62	4.08 / 2.68	1.68 / 1.93	3.28 / 2.32
8	5.87 / 3.73	5.22 / 3.35	5.61 / 4.11	4.71 / 3.16
9	6.57 / 4.08	5.22 / 3.13	5.14 / 3.63	4.90 / 2.92
10	5.02 / 4.00	6.08 / 5.06	4.59 / 3.52	5.71 / 4.71
Ave.	4.74 / 3.88	5.15 / 3.97	3.74 / 3.42	4.70 / 3.71



(a)



(b)

Fig. 11: Average SDR of 10 test songs with different objective functions defined by varying β. The thickest (black) lines, the second thickest (red) lines, and the thinest (blue) lines represent ST-DSS, S-DSS, and T-DSS results, respectively. Dotted, solid, and dashed lines represent when $\beta = 0$, 1, and 2, respectively. In each experiment, parameters were set to the best ones for the input test song 1. (a) 2048 FFT (b) 4096 FFT

60th iteration in this case. The experimental results in TABLE III show that the average SDR of ST-DSS degrades slower than S-DSS results. For instance, while S-DSS degrades by 1 and 0.46 dB in the case of $\beta = 1$ and 2, respectively, ST-DSS is lowered only by 0.45 and 0.26. S-DSS closes these gaps in 4096 FFT size tests in TABLE IV, but the tendency that ST-DSS is less sensitive to the random choice of stopping time, the number of iterations, still holds.

The authors also conducted Wiener filtering for above all experiments to reconstruct time domain signals. Since the results contained less artifacts caused by reusing mixture phase, Wiener filtering can be a good way to reconstruct final results although they usually contain a little more interference and do not provide significantly higher SDR.

V. Discussion and Conclusion

In this work, a unified approach to drum source separation was proposed. The proposed method tried to improve two NMPCF-based predecessors by harmonizing their heterogenious assumptions: there are distinct features of drum sources both in the spectral and temporal domains. The main contribution of this paper is that it provides adequate formation of input signals, by grouping them into prior knowledge and partitioned column-blocks of mixture signals, which are allocated delicately to NMPCF learning process. The experimental results showed advantages of the proposed ST-DSS system with real-world music signals. ST-DSS mechanism successfully incorporated two previous NMPCF-based predecessors, by both exploiting spectral bases of a priori drum solo signals and native temporal structures of drum sources. Also, it generally improved the low separation performance of the T-DSS method while retaining its robustness to diverse parameter configuration, which is a drawback of the S-DSS scheme.

As spectrogram factorization based recent source separation methods employed, we have also tried to assess various divergence criteria, such as KL-divergence and IS-divergence, but additional massive parameter investigation that might be needed prevented meaningful comparison of those divergences. Furthermore, we checked that higher frequency resolution does provide better separation of bass drum sources from bass guitars for specific input mixtures, but enlarged FFT size did not generally improve separation performance because it also requires further investigation of parameters.

In the future, the authors want to research the relationship between the size of prior-information matrix for S-DSS or ST-DSS and separation performance. For now, it is known that too long drum solo auxiliary input does not improve the separation quality, because incongruent parts can harm the reconstruction. Also, the authors are also concerned with applying the proposed method to the other MSS tasks. Another possible imperfection of this work is that conventional objective assessment using SDR is not sufficient to judge the sound quality of the poorly isolated drum sources in single channel MSS tasks. Along with subjective tests, recent suggestions about objective measurements of audio quality which take

TABLE IV: Separation performances of S-DSS and ST-DSS for the 10 commercial songs. Same setting with experiments in TABLE III was used, except that FFT size was set to 4096.

	iteration=20		iteration=60	
	S-DSS	ST-DSS	S-DSS	ST-DSS
Song	β=1 / β=2	β=1 / β=2	β=1 / β=2	β=1 / β=2
1	4.81 / 4.84	2.98 / 5.70	3.68 / 4.90	3.43 / 5.97
2	4.66 / 5.00	5.92 / 5.62	3.98 / 4.22	6.35 / 5.87
3	3.94 / 3.73	3.69 / 3.90	3.66 / 3.59	3.53 / 3.87
4	4.12 / 4.50	2.66 / 5.87	3.47 / 4.43	2.60 / 6.11
5	4.34 / 1.39	4.36 / 4.46	4.68 / 1.60	3.67 / 4.03
6	6.69 / 2.73	2.98 / 6.63	6.22 / 2.82	2.60 / 6.65
7	5.25 / 5.44	6.50 / 5.89	4.21 / 4.94	6.55 / 6.32
8	5.74 / 2.11	5.79 / 4.56	5.82 / 2.37	5.26 / 4.61
9	5.65 / 3.74	3.88 / 5.30	5.88 / 3.91	4.06 / 5.34
10	3.91 / 3.61	3.53 / 3.93	3.93 / 3.59	3.54 / 3.71
Ave.	4.91 / 3.71	4.23 / 5.19	4.55 / 3.64	4.16 / 5.25

properties of human auditory perception into account [32] [33] are expected to provide more perceptively convincible comparison.

REFERENCES

[1] I. Jang, J. Seo, and K. Kang, "Design of a file format for interactive music service," *ETRI Journal*, vol. 33, no. 1, pp. 128–131, 2011.

[2] *Information technology – Multimedia application format (MPEG-A) – Part12: Interactive music application format*, ISO/IEC IS 23 000-12, 2010.

[3] E. Tsunoo, N. Ono, and S. Sagayama, "Musical bass-line pattern clustering and its application to audio genre classification," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Kobe, Japan, 2009.

[4] E. Tsunoo, Taichi, Akase, N. Ono, Shigeki, and Sagayama, "Music mood classification by rhythm and bass-line unit pattern analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[5] P. Chordia and A. Rae, "Using source separation to improve tempo detection," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Kobe, Japan, 2009.

[6] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Neural Information Processing Systems Workshop on Advances in Models for Acoustic Processing*, 2006.

[7] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 529–540, 2008.

[8] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2008, pp. 139–144.

[9] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of EUSIPCO*, 2008.

[10] D. FitzGerald, E. Coyle, and M. Cranitch, "Using tensor factorisation models to separate drums from polyphonic music," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Como, Italy, 2009.

[11] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Nara, Japan, 2003.

[12] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *European Signal Processing Conference*, 2005.

[13] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

[14] M. Kim, J. Yoo, K. Kang, and S. Choi, "Blind rhythmic source separation: Nonnegativity and repeatability," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

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

[16] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 13. MIT Press, 2001.

[17] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[18] M. Kim and S. Choi, "On spectral basis selection for single channel polyphonic music separation," in *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, vol. 2. Warsaw, Poland: Springer, 2005, pp. 157–162.

[19] ——, "Monaural music source separation: Nonnegativity, sparseness, and shift-invariance," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*. Charleston, South Carolina: Springer, 2006, pp. 617–624.

[20] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted nonnegative matrix factorisation for sound source separation," in *IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, 2005.

[21] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, 2005.

[22] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining content and link for classification using matrix factorization," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Amsterdam, The Netherlands, 2007.

[23] H. Lee and S. Choi, "Group nonnegative matrix factorization for EEG classification," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, Florida, 2009.

[24] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Las Vegas, Nevada, 2008.

[25] J. Yoo and S. Choi, "Weighted nonnegative matrix co-tri-factorization for collaborative prediction," in *Proceedings of 1st Asian Conference on Machine Learning*, Nanjing, China, 2009.

[26] M. Minami and S. Eguchi, "Robust blind source separation by beta-divergence," *Neural Computation*, vol. 14, pp. 1859–1886, 2002.

[27] A. Cichocki and S. Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 2010, no. 12, pp. 1532–1568, 2010.

[28] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[29] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[30] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, USA, 2010.

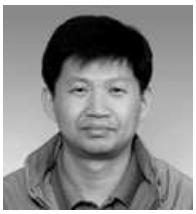
[31] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[32] M. Lee, I. Heo, N. Choi, and K. Sung, "On evaluation of blind audio source separation," in *Audio Engineering Society 34th International Conference*, Jeju, Korea, 2008.

[33] T. Kastner, "Evaluating physical measures for predicting the perceived quality of blindly separated audio source signals," in *Audio Engineering Society 127th Convention*, New York, 2009.

Minje Kim received his BS degree in Information and Computer Engineering from Ajou University, Suwon, Korea, in 2004, and his MS degree in Computer Science from Pohang University of Science and Technology, Pohang, Korea, in 2006, respectively. He is currently a member of research staff in Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. He is interested in the music signal processing, blind source separation and audio coding using machine learning techniques.

Jiho Yoo received BS in Computer Science and Mathematics from Pohang University of Science and Technology. Currently he is a PhD candidate at department of computer science in Pohang University of Science and Technology. He is interested in the nonnegative matrix factorizations and its applications to document clustering, collaborative prediction, and musical source separation.



Kyeongok Kang received his BS and MS degrees in physics from Pusan National University, Busan, Korea, in 1985 and 1988, respectively, and his PhD degree in electrical engineering from Hankuk Aviation University, Seoul, Korea, in 2004. He has been with ETRI since 1991, and he is now a principal member of engineering staff and the leader of the Realistic Acoustics Research Team. His major interests are in low-bitrate audio coding; audio signal processing, including 3-dimensional audio and personalized broadcasting based on MPEG-7; and TV-Anytime related issues.

Seungjin Choi received the B.S. and M.S. degrees in Electrical Engineering from Seoul National University, Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical engineering from the University of Notre Dame, Indiana, in 1996. He was a Visiting Assistant Professor in the Department of Electrical Engineering at University of Notre Dame, Indiana, during the Fall semester of 1996. He was with the Laboratory for Artificial Brain Systems, RIKEN, Japan, in 1997 and was an Assistant Professor in the School of Electrical and Electronics Engineering, Chungbuk National University from 1997 to 2000. He is currently a Professor of Computer Science at Pohang University of Science and Technology, Korea. His primary research interests include machine learning, Bayesian inference, and probabilistic models.