Dual-channel Drum Separation for Low-cost Drum Recording Using Non-negative Matrix Factorization

Cheng-Yu Cai,* Yu-Hui Su‡ and Li Su†
*‡Research Center of Music, Technology and Health, National Tsing Hua University, Hsinchu, Taiwan
†Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: *s107591505@m107.nthu.edu.tw, ‡yhsu@mail.nd.nthu.edu.tw, †lisu@iis.sinica.edu.tw

***Abstract*—Recording drum performance is more complicated than recording other music performance because a drum set contains multiple classes of percussive instruments each of which cannot be recorded individually. In this work, we consider a low-cost drum recording setup, which adopts dual-channel source separation techniques based on the non-negative matrix factorization (NMF) and non-negative matrix factor deconvolution (NMFD) techniques to reconstruct the eight percussive instrument tracks from recordings with only two overhead microphones. Compared to previous drum separation works which only separate bass drum, snare drum and hi-hat, this work firstly considers the separation of tom drums and cymbals, which is a technically challenging task. A pipeline for imitating the panning of each instrument of a target recording is also designed. Results of source separation and panning imitation demonstrate the potential of the proposed method on low-cost drum recording.**

I. Introduction

Human's hearing system is capable of processing complex information. Suppose we are listening to a symphony, we hear the violin, trumpet and percussion at the same time, and we can distinguish each sound from others with our ability called *cocktail party effect*. However, this does not mean that the each sound source in such an environment can be easily reproduced. How to record clean sound sources in noisy environments is therefore a long-lasting challenge of recording engineering.

One of the solutions commonly adopted among professionists is using multiple microphones with delicately adjusted directions to reduce the *crosstalk* between sound sources during recording, which is however an expensive and less portable solution. To reduce the cost, an alternative solution is *multitrack recording*, which allows each instrument track to be recorded separately. These tracks are then assigned to the mixing engineer for post-production. Since home studio recording has become prevailing in recent years, multitrack recording is gaining its popularity among the communities. Performing well for most instruments, multitrack recording is however not applicable for drum set recording: it is impossible to record each percussion instrument separately in a natural drum performance. An effective, low-cost, and personalized drum set recording framework allowing separated drum track recording in home studio is therefore a valuable research topic.

In this work, we attempt to provide a solution to this research problem using audio source separation techniques to obtain clean sources from mixture recordings, such that the recording process can be done with simple equipment such as overhead microphones. Most research on audio source separation are aimed at denoising (i.e. removing background noise) [1], vocal separation (i.e. separating background music and human voice) [2], or harmonic-percussive source separation (HPSS, i.e. separating percussive and pitched instruments) [3]. Research on *drum separation*, i.e. separating individual percussive instrument in a drum set is relatively less discussed [4], [5]. A standard drum set contains eight classes of instruments, but all the previous work on drum separation only discussed the separation of the three most significant instruments (bass drum, snare drum, and hi-hat).

Compared to other types of audio source separation tasks, the major difficulty of drum separation is that the spectra of individual instruments in the drum set are highly overlapped in both frequency and time: drum signals are typically wideband and drum events tend to occur at beat or sub-beat positions. For example, bass drum and floor tom share similar frequency band and snare drum are usually played together with hi-hat. Previous drum separation methods are mostly based on the non-negative matrix factorization (NMF) and non-negative matrix factor deconvolution (NMFD) algorithm, which in principle takes the spectrogram of the drum set performance as the input and decompose it into a *template* matrix representing the spectral pattern of each percussive instrument and an *activation* matrix representing the active time of each of them (see Section III-A). In comparison to the most recently developed deep-learning-based approach [6], the NMF-based approach is still advantageous to our scenario because the outcomes of the algorithm (i.e. templates and the activation footprints) are all interpretable and can facilitate the downstreaming tasks in recording engineering, such as automatic adjustment of panning level (i.e., the panning imitation task, see Section III C).

We set our research problem as: separating eight percussive instrument sources (see Section III-B) from merely a mixture recording of a pair of overhead microphone using the NMF-based algorithm, such that the quality and format of the separated results suffice further sound reproduction tasks. The major contributions of this paper are two-fold:

- To the best of our knowledge, this work represents the first attempt to solve drum set separation of problem with the all the drum set instruments.
- This work is also the first discussion on applying drum separation techniques for low-cost drum recording.

The source code and detailed experiment results can be found on https://github.com/aaron985/dual_channel-NMF-for-drum-separation.

II. Background

A. Drum set and drum recording techniques

A drum set is in general constructed with *membranophone* including bass drum (BD), snare drum (SD), tom-tom 1 (T1), tom-tom 2 (T2), and floor drum (FT), as well as *idiophone* including hi-hat (HH), crash cymbal (CC) and ride cymbal (RC). The bass drum takes the largest room and produces the lowest sound. The snare drum sounds sharp with the attached snare. The hi-hat consists in two face-to-face cymbals and can produce short and high sound when closing them up. Tom-tom 1, tom-tom 2 and floor drum are drums of different sizes and similar appearances. Crash cymbal and ride cymbal are characterized with the long sustain sounds.

The standard drum set recording method requires eleven microphones. The standard setting are described as follows. Two microphones are needed for bass drum: one is placed in front of the resonance drumhead, and the other is placed inside the bass drum near the beat drumhead. Two microphones are needed for snare drum: one is placed on the beat drumhead above the snare drum; the other is placed on the resonance drumhead below the snare drum. One microphone is need on the top of the hi-hat. The tom-tom 1, tom-tom 2 and floor tom each need one microphone placed above the beat drumhead. Ride cymbal needs one microphone placed above the cymbal and aimed at the center of the cymbal. Lastly, two symmetrical microphones are needed for crash cymbal and entire drum set. Two symmetrical microphones are called "overhead microphone" (OH) are placed symmetrically on the on right and left side above the drum set with the entire drum set as the center. In some situations, the crash cymbal is recorded with overhead microphones. In the practice for drum recording, using only the two OH microphones is also a commonly-seen, simplified setting. The OH microphones can receive the sounds of all instruments of a complete drum set. It is then intriguing to see if we can apply source separation techniques on the two OH signals to approximate the above-mentioned standard setting.

B. Related work on drum separation

Audio source separation is a classic research field in signal processing. In this section, our survey focuses specifically on the drum separation problem. Survey on general audio source separation can be found in [7]. To our knowledge, there is still no specific review paper on drum separation, though interested readers can refer to the review of its most related task, automatic drum transcription [8].

Most of the drum spearation methods are based on spectorgram decomposition. Yoo *et al.* [9] proposed to use Non-negative Matrix Partial Co-factorization (NMPCF) to separate the drum sound from the melody instrument sound in the music signal. In this method, a pre-trained classifier that distinguishes the drum sound from the melody instrument sound is not required as it can decompose the feature of percussive and non-percussive instruments. Dittmar *et al.* [4] further utilized the high-efficiency of NMF to separate the sources of three percussive instrument (i.e. hi-hat, snare drum, and bass drum) in a drum set in real-time and transcribe them into symbolic form. Besides, Dittmar *et al.* [5] also considered score-informed drum separation, which was based on score-informed NMFD. It requires both the training data of drum signals and the note-level annotations (i.e. scores) of the testing data. The score information can guide the learning of the activation matrix and improve the performance. Rathnayake [10] proposed to use NMF to find the optimal dictionary atoms for monophonic source separation based on the Local Non-negative Matrix Factorization (LNMF) method and stability-driven model selection criterion. Pachauri *et al.* [11] proposed to use NMF to separate beatbox, the art that uses human voices to simulate various percussion sounds and rhythms. There are three sets of sound data in the experiment. It can be seen from the literature that NMF is a feasible method to separate the drum sounds in an efficient and interpretable manner with small-scale training data.

C. NMF-based audio reproduction

Using signal processing and optimization techniques on various audio reproduction tasks such as automatic audio mixing [12], demixing [13], and conversion [14], [15] has been widely seen in the literature. The main applications of these techniques include generating novel audio effects and reducing cost in audio recording and post-production.

Scott *et al.* [16] proposed to use the linear dynamic system to automatically mix the music stem. The part of the drum set uses the Non-negative Least Squares (NNLS) method to calculate the weight of each instrument in the drum set, and then multiplies the frequency spectrum of each drum set by the weight and mixes them all together to get the automatic mixing result of the drum set. This automatic mixing method mainly performs for the volume part rather than the stereo parameters such as panning. Su [14] proposed a system to automatically convert pop music into 8-bit chiptune music. It adopted Robust Principal Component Analysis (RPCA) to separate the vocals and musical instruments in pop music, and NMF to transcribe the activation of musical instruments. By replacing the template matrix of musical instruments with chiptune-like templates, the chiptune music can be synthesized. Such an NMF-based audio conversion framework is widely adopted in various scenarios [15].

III. Proposed Method

A. NMF and NMFD

The NMF algorithm decompose a mixture spectrogram $\mathbf{A} \in \mathbb{R}_{\geq 0}^{F \times T}$ into a template matrix (or $\mathbf{W}$ matrix for simplicity) $\mathbf{W} \in \mathbb{R}_{\geq 0}^{F \times R}$ and an activation matrix (or $\mathbf{H}$ matrix for simplicity) $\mathbf{H} \in \mathbb{R}_{\geq 0}^{R \times T}$ by finding a low-rank approximation of $\mathbf{A}$ such that $\mathbf{A} \approx \hat{\mathbf{W}}\mathbf{H}$, under the condition that $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$ are all non-negative matrices. The values F, T, and R are the number of frequency bins, the number of time frames and the

number of templates, respectively. The objective function of the NMF algorithm is to minimize $D(\mathbf{A}|\mathbf{WH})$, where $D(\cdot|\cdot)$ represents the distance of two matrices. Conventionally used distance functions in NMF include the Euclidean distance, the Kullback-Leibler (KL) divergence, and the Itakura-Saito (IS) divergence, etc. We adopt Euclidean distnace in this work as it is found more stable in drum separation based on our pilot study. The NMF algorithm is iterative with $\mathbf{W}$ and $\mathbf{H}$ updated alternatively, and each step of update contains only matrix multiplication and some element-wise operation. Details of the algorithm can be found in [17].

The NMFD algorithm is an extension of NMF by introducing a convolution between a template and the activation sequence over the time axis. That means, each template in NMFD is two-dimensional and the whole set of templates aggregate into a *template tensor* $\mathbf{P} \in \mathbb{R}_{\geq 0}^{F \times R \times C}$, in which C is the number of feature frames each two-dimensional template. Denote $\mathbf{P}_c$ as a slice of $\mathbf{P}$ at the cth frame of the template, $\mathbf{A}$ is decomposed with NMFD in the following way:

$$\mathbf{A} \approx \sum_{c=1}^{C-1} \mathbf{P}_c \overrightarrow{\mathbf{H}}^{(c)}, \tag{1}$$

where C is the total number of template and $\overrightarrow{\mathbf{H}}^{(c)}$ is the frame shift of $\mathbf{H}$ by c frames from left to the right [18], [19]. With the temporal convolution, NMFD is found to be better in learning the temporal behaviors of the signal. Details of the NMFD algorithm can be found in [18], [19].

B. Dual-channel drum separation

Figure 1 illustrates the proposed dual-channel drum separation framework. In our framework, we consider using NMFD for drum separation, and its initial template and activation matrices are estimated using NMF over the magnitude spectrogram of the training data. Given a dual-channel drum recording $(\mathbf{a}_L, \mathbf{a}_R)$ which are recorded by the left overhead microphone (OHL) and the right overhead microphone (OHR), respectively. Denote the spectrogram of $(\mathbf{a}_L, \mathbf{a}_R)$ as $(\mathbf{A}_L, \mathbf{A}_R)$. The drum spearation module outputs eight signals for each channel, denoted as $(\bar{\mathbf{a}}_{L,i}, \bar{\mathbf{a}}_{R,i})$, such that

$$(\mathbf{a}_L, \mathbf{a}_R) \approx \sum_{i \in \mathcal{I}} (\bar{\mathbf{a}}_{L,i}, \bar{\mathbf{a}}_{R,i}) \tag{2}$$

where the set of subscripts $\mathcal{I} := \{\mathrm{BD}, \mathrm{SD}, \mathrm{HH}, \mathrm{T1}, \mathrm{T2}, \mathrm{FT}, \mathrm{CC}, \mathrm{RC}\}$ represent the eight percussive instrument classes. The proposed dual-channel NMF process aims at computing the individual template and activation matrices of the two signals, and also finding the relation between them. The source separation process includes the following steps:

1) **Template initialization.** The spectra of each instruments in the training dataset are selected and concatenated as the initial template matrix $\mathbf{W}_0 := [\mathbf{W}_{0,\mathrm{BD}}|\mathbf{W}_{0,\mathrm{SD}}|\cdots|\mathbf{W}_{0,\mathrm{RC}}]$. $\mathbf{W}_0$ is used as the initial template matrix in the following NMF process.
2) **Individual-channel decomposition using NMF.** $\mathbf{A}_L$ and $\mathbf{A}_R$ are decomposed using NMF such that $\mathbf{A}_L \approx \mathbf{W}_L\mathbf{H}_L$ and $\mathbf{A}_R \approx \mathbf{W}_R\mathbf{H}_R$, with $D(\mathbf{A}_L|\mathbf{W}_L\mathbf{H}_L)$ and $D(\mathbf{A}_R|\mathbf{W}_R\mathbf{H}_R)$ minimized. $\mathbf{W}_L$, $\mathbf{W}_R$, $\mathbf{H}_L$, and $\mathbf{H}_R$ are then the template and activation matrices for either the left or the right channels according to the subscript.
3) **Cross-channel decomposition using NMF.** Assuming that there exists time and phase difference between each instrument to the OHL and OHR microphones, a *communication matrix* $\mathbf{U}$ between the left and right channels is defined to communicate the left and right channels. Initializing $\mathbf{U}$ with an identity matrix, two NMF processes are then performed: first, update $\mathbf{W}_L$ and $\mathbf{U}$ by $\mathbf{W}_R \approx \mathbf{W}_L\mathbf{U}$; second, update $\mathbf{W}_R$ and $\mathbf{U}$ by $\mathbf{W}_L \approx \mathbf{W}_R\mathbf{U}$. With the updated $\mathbf{W}_R$ and $\mathbf{W}_L$ in the above step, update $\mathbf{W}_L$, $\mathbf{W}_R$, $\mathbf{H}_L$, and $\mathbf{H}_R$ again.
4) **Source separation with NMFD.** $\mathbf{A}$ is then decomposed again with NMFD using the $\mathbf{W}$ and $\mathbf{H}$ obtained from the above step as initialized template and activation matrices. The template tensor $\mathbf{P}$ and the activation matrix $\mathbf{H}$ obtained in this step are tensors containing 8 channels.
5) **Reconstruction.** The spectrogram for the instrument i, denoted as $(\bar{\mathbf{A}}_{L,i}, \bar{\mathbf{A}}_{R,i})$, is reconstructed by this formulation: $\bar{\mathbf{A}}_i := \sum_{c=0}^{C-1} \mathbf{P}_{c,i} \overrightarrow{\mathbf{H}}^{(c)}$ (subscripts L and R are omitted), in which $\mathbf{P}_{c,i}$ is denoted as the template sets for instrument i. Each $\bar{\mathbf{A}}_i$ in both channels is processed with an alpha Wiener filter [20], and is then converted

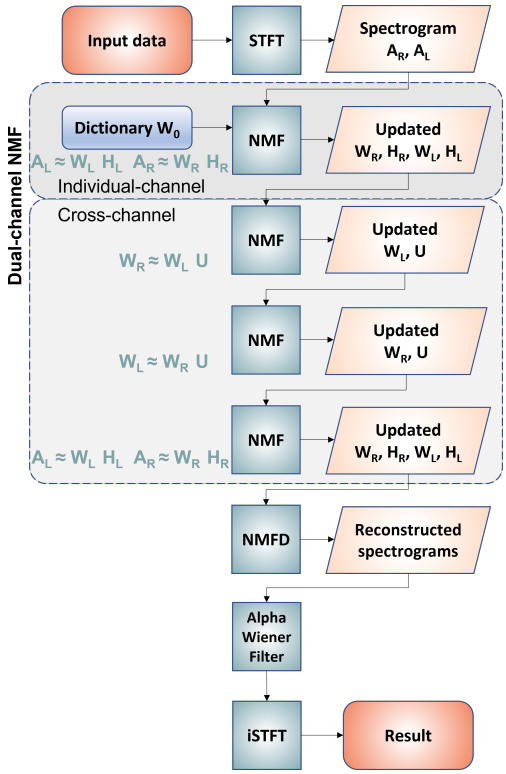

Fig. 1. Flowchart of the dual-channel drum separation system.

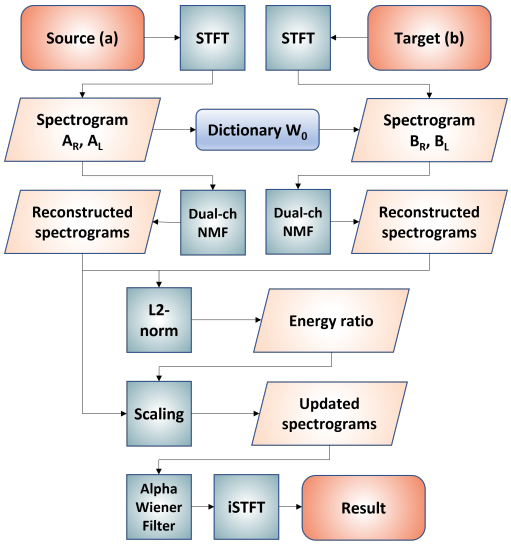

Fig. 2. Flowchart of the panning imitation system.

back to the time-domain signal $\{\bar{\mathbf{a}}_{L,i}, \bar{\mathbf{a}}_{R,i}\}_{i\in\mathcal{I}}$ with an inverse short-time Fourier transform (iSTFT) process implemented with inverse fast Fourier transform (IFFT) and the overlap-add (OLA) technique [21]. In the iSTFT process, the phase spectrogram of the source signal is used for reconstruction.

C. Panning Imitation

Briefly speaking, panning refers to the ratio of levels between the left channel and the right channel. Panning imitation is the task to estimate the panning level of each instrument in the input signal (i.e. the *source signal*) as well as in a well-recorded drum set signal (i.e. the *target signal*), such that the panning condition of the input can be converted to the well-recorded one. Panning imitation is an extra application that applies the dual-channel drum separation technique for low-cost drum recording. In this task, the source signal is denoted as $\mathbf{a} := (\mathbf{a}_L, \mathbf{a}_R)$ and the target signal as $\mathbf{b} := (\mathbf{b}_L, \mathbf{b}_R)$. Panning imitation is done with the following processes:

1) Both the source and the target signals are separated into $\{\mathbf{a}_{L,i}, \mathbf{a}_{R,i}\}_{i\in\mathcal{I}}$ and $\{\mathbf{b}_{L,i}, \mathbf{b}_{R,i}\}_{i\in\mathcal{I}}$, respectively, using the proposed dual-channel drum separation technique. It should be noted that the initialized $\mathbf{W}$ and $\mathbf{H}$ matrices for separating $\mathbf{b}$ are obtained from the outcomes when separating $\mathbf{a}$. Since the panning information of the source is not necessary for panning imitation, $\mathbf{a}_{L,i}$ and $\mathbf{a}_{R,i}$ are merge into a single-channel signal (denoted as $\mathbf{a}_i$) simply by taking average.
2) The average energy values of the each instrument track after source separation are obtained by performing l_2-norm for each spectrogram. These values are denoted as $\{A_i\}_{i\in\mathcal{I}}$, $\{B_{L,i}\}_{i\in\mathcal{I}}$, and $\{B_{R,i}\}_{i\in\mathcal{I}}$ for the source signal, left-channel target signal and right-channel target signal, respectively.

TABLE I
TRAINING DATA

Class	Techniques	#. file	Length (s)
BD	center	13	78
SD	center. rim shot, side stick	30	180
HH	closed edge, half closed edge, half open edge	27	162
T1	center	9	54
T2	center	8	54
FT	center	12	72
CC	edge	51	306
RC	edge, bell	22	132

3) The imitation results, $(\bar{\mathbf{a}}_L, \bar{\mathbf{a}}_R)$, are then

$$\bar{\mathbf{a}}_L = \sum_{i\in\mathcal{I}} \frac{B_{L,i}}{A_i}\mathbf{a}_i\,, \quad \bar{\mathbf{a}}_R = \sum_{i\in\mathcal{I}} \frac{B_{R,i}}{A_i}\mathbf{a}_i\,. \tag{3}$$

D. Implementation details

Throughout this work, the audio signals are sampled with a rate of 44.1 kHz and a bit depth of 16 bits. The parameters for STFT computation are: Hann window function, 256-point hop size, and 2048-point frame size.

The code was written in MATLAB, and the implementation mostly relied on the NMF Toolbox [19].[1] The distance function used in this work is Euclidean distance for NMF and KL divergence for NMFD. Each NMF/NMFD process mentioned in this paper is run with 15 iterations. If not specifically mentioned, the $\mathbf{W}$ matrix is randomly initialized and the $\mathbf{H}$ is uniformly initialized (i.e. all elements are one) for NMF. For NMFD, the $\mathbf{P}$ matrix is initialized with the `drums` strategy in `initTemplates.m`, which converts the 2D $\mathbf{W}$ matrix to a tensor. The number of template R is 172, and the number of frames of a two-dimensional template C is 64. The values smaller than 1×10^{-8} in $\mathbf{H}$ are discarded in order to sparsify the $\mathbf{H}$ matrix.

IV. Experiment and Results

A. Data

The training data for each instrument are extracted from the single-strike samples provided by Perfect Drums, an acoustic drum VST.[2] The extraction process from MIDI to audio is performed on Cubase 10 Pro.[3] The MIDI velocity of each sample is set to 80. Each audio file is a mono-channel .wav file with a bit depth of 16 bits and a sampling rate of 44.1 kHz. The training dataset therefore contains the aforementioned eight percussive instruments with various techniques (e.g., tapping positions, hi-hat modifiers), and the total length of the dataset is 1038 seconds. See Table I for details.

The testing data are selected from the OHL/OHR samples as well as the ground truth signals provided by the ENST Drums Dataset [22].[4] The ENST Drums Dataset includes the recordings of three drummers. Among them, drummer 1's

[1] https://www.audiolabs-erlangen.de/resources/MIR/NMFtoolbox/
[2] https://theperfectdrums.com/
[3] https://new.steinberg.net/cubase/
[4] https://perso.telecom-paristech.fr/grichard/ENST-drums/

TABLE II
TESTING DATA

	#. file	Length (s)	Name
drummer 2	3	143	072_phrase_shuffle-blues_complex_slow_sticks 079_phrase_hard-rock_complex_medium_sticks 106_solo_binary_sticks
drummer 3	2	145	087_phrase_shuffle-blues_simple_fast_sticks 089_phrase_shuffle-blues_complex_medium_sticks

recordings are note muffed and represents a more challenging scenario for source separation due to the lengthened sustain of the sound. Therefore, we opt to experiment on selected recordings performed by drummer 2 and drummer 3. We selected the recordings with all the eight drum set instruments and without those percussive instruments not discussed in this paper (e.g., cowbell). This results in five recordings, as listed in Table II.

B. Experiment settings and evaluation metrics

To verify the effectiveness of the proposed dual-channel drum separation method, we consider two baseline methods in our experiment. The first baseline is the proposed method with the cross-channel decomposition step removed, which is then equivalent to the typical, single-channel NMF and NMFD process for drum separation (we will refer to this setting as single-channel NMF hereafter). In this way we can evaluate whether the method takes advantage of using the dual-channel setting. The second baseline is Drums SSX , which is a commercial tool for separating drum sounds based on NMF.[5]

We use the BSS Eval 3.0 toolkit to evaluate the performance of source separation and panning imitation. The source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR) are reported. All the reported values are in decibel (dB). It should be noted that in our testing data, there is no ground truth signals for the crash cymbal and the ride cymbal as there is no specific microphone devices used for recording these two instruments (in practice, these two instruments are recorded with the OH microphones only).

For panning imitation, there is also no available ground truth (i.e. the left/right-channel energy distribution of each instrument in the target signal) since the right- and left-channel of the target are both mixed. We therefore consider using the SDR, SAR and SIR between the target $\mathbf{b}$ and the remixed source after panning imitation $\bar{\mathbf{a}}$ as a preliminary measure, by assuming that *these performance metrics for the results with an effective panning imitation process should be higher than those without panning imitation.* More specifically, six values are for comparison: $\mathrm{SDR}(\bar{\mathbf{a}}_L, \mathbf{b}_L)$, $\mathrm{SDR}(\bar{\mathbf{a}}_R, \mathbf{b}_R)$, $\mathrm{SIR}(\bar{\mathbf{a}}_L, \mathbf{b}_L)$, $\mathrm{SIR}(\bar{\mathbf{a}}_R, \mathbf{b}_R)$, $\mathrm{SAR}(\bar{\mathbf{a}}_L, \mathbf{b}_L)$, and $\mathrm{SAR}(\bar{\mathbf{a}}_R, \mathbf{b}_R)$. The higher these values are, the better the system performs. Three scenarios are considered in order to verify the effectiveness of panning imitation and its performance on various targets:

1) $(\bar{\mathbf{a}}_L, \bar{\mathbf{a}}_R)$ are obtained with panning imitation, and the target is the *wet mix* version of that source recording which is provided in the ENST-Drum dataset.
2) $(\bar{\mathbf{a}}_L, \bar{\mathbf{a}}_R)$ are obtained with panning imitation, and the target is an artificially-mixed OH recording which is simply mixed by $0.7 \times .\mathrm{OHR} + 0.3 \times \mathrm{OHL}$.
3) The remixed sources $(\bar{\mathbf{a}}_L, \bar{\mathbf{a}}_R)$ are obtained without panning imitation (i.e. $B_{L,i}/A_i = B_{R,i}/A_i = 0.5$ for all i), and the target is the *wet mix* version of that source recording.

[5]Drum SSX: https://fuseaudiolabs.com/#/pages/product?id=300867907

C. Results

Table III lists the SDR, SIR and the SAR values on the testing dataset for the proposed dual-channel drum separation method, the proposed method without cross-channel decomposition (i.e., single-channel), and Drum-SSX. Again, the performance of CC and RC cannot be reported due to the lack of ground truth signals. The optimal values among the three methods are marked in bold. First, both the single- and dual-channel NMF greatly outperforms Drum-SSX: taking SDR for example, dual-channel NMF leads Drum-SSX by around 10 – 20 dB for each instrument. This indicates the effectiveness of the proposed method in drum separation. Second, the single-channel and dual-channel methods achieve similar performance, though dual-channel NMF still outperform single-channel NMF quite consistently in a range between 0.1 and 0.5 dB. Single-channel NMF outperforms dual-channel NMF only in the SIR of snare drum and hi-hat. In summary, the cross-channel decomposition process in dual-channel NMF results in small but consistent improvement in drum separation.

Table IV shows the SDR, SAR and SIR between the target signal (wet mix or OH) and the remixed signal with or without panning imitation. Results show that in all cases, panning imitation results in better performance. It should be emphasize that even in the very challenging case (i.e. wet mix as the target), the performance metrics of the left and right channel are still consistently better than the case without panning invitation. Finally, the results that the system can imitate OH better than wet mix are also reasonable.

V. CONCLUSION

We have demonstrated a drum separation system which can separate eight percussive instrument in a drum set from either single-channel or dual-channel inputs. Based on this drum separation system, we have also demonstrate its practical use of imitating the panning levels of a target recording. Both of them are of high potential in the re-production of low-cost drum recordings. However, the performance of drum separation is still strongly limited by several factors, such as the high similarity in between the instruments and the paucity

TABLE III
RESULTS OF SDR, SIR, AND SAR (IN DB) FOR THE DRUM SOURCE SEPARATION METHODS

	dual-channel NMF			single-channel NMF			Drums SSX		
	SDR	SAR	SIR	SDR	SAR	SIR	SDR	SAR	SIR
BD	**-3.549**	**-1.292**	**4.226**	-3.779	-1.457	3.972	-15.52	-12.213	-0.029
SD	**6.975**	**10.249**	10.323	6.926	10.107	**10.422**	-21.827	-11.535	-9.523
HH	**-10.673**	**-6.470**	-1.218	-10.748	-6.904	**-0.714**	-19.352	-13.602	-4.046
T1	**-2.470**	**4.217**	**0.760**	-2.835	4.05	0.402	-16.692	-12.24	-1.465
T2	**-2.827**	**2.288**	**0.928**	-3.089	2.155	0.681	-15.148	-11.707	0.220
FT	**-5.508**	**0.658**	**-1.591**	-5.754	0.456	-1.758	-17.893	-13.792	-1.723

TABLE IV
RESULT OF PANNING IMITATION

Panning imitation	Yes	Yes	No
Target (**b**)	Wet mix	Overhead	Wet mix
SDR (right)	-5.892	-2.606	-6.589
SDR (left)	-6.186	-2.706	-6.760
SIR (right)	2.890	7.095	2.905
SIR (left)	2.757	6.619	2.723
SAR (right)	-3.391	-1.283	-4.236
SAR (left)	-3.536	-1.194	-4.236

of training data and full (8-channel) ground truth. We also showed that the limited performance of drum separation is still the major obstacle of deploying it to the downstream tasks of audio reproduction. The task of collecting scaled multi-track drum recording data would then be the most essential future work. Combining NMF with deep learning techniques for drum separation is also a provoking research direction [23].

REFERENCES

[1] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2267–2282, 2020. I

[2] A. Jansson, R. M. Bittner, S. Ewert, and T. Weyde, "Joint singing voice separation and f0 estimation with deep u-net architectures," pp. 1–5. I

[3] Y.-N. Hung and A. Lerch, "Multitask learning for instrument activation aware music source separation," *arXiv preprint arXiv:2008.00616*, 2020. I

[4] C. Dittmar and D. Gärtner, "Real-time transcription and separation of drum recordings based on nmf decomposition," pp. 187–194. I, II-B

[5] C. Dittmar and M. Müller, "Reverse engineering the amen break — score-informed separation and restoration applied to drum recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1535–1547, 2016. I, II-B

[6] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," pp. 258–266, 2017. I

[7] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018. II-B

[8] C.-W. Wu, C. Dittmar, C. Southall, R. Vogl, G. Widmer, J. Hockman, M. Müller, and A. Lerch, "A review of automatic drum transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1457–1483, 2018. II-B

[9] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for drum source separation," pp. 1942–1945. II-B

[10] B. Rathnayake, K. M. K. Weerakoon, G. M. R. I. Godaliyadda, and M. P. B. Ekanayake, "Toward finding optimal source dictionaries for single channel music source separation using nonnegative matrix factorization," pp. 1493–1500. II-B

[11] N. Pachauri and M. Wajid, "Single channel beatbox music separation using non-negative matrix factorisation," pp. 1011–1017. II-B

[12] B. De Man, J. Reiss, and R. Stables, "Ten years of automatic mixing," 2017. II-C

[13] D. Fitzgerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1560–1572, 2016. II-C

[14] S. Su, C. Chiu, L. Su, and Y. Yang, "Automatic conversion of pop music into chiptunes for 8-bit pixel art," pp. 411–415. II-C, II-C

[15] H. F. Aarabi and G. Peeters, "Music retiler: Using nmf2d source separation for audio mosaicing," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, pp. 1–7, 2018. II-C, II-C

[16] J. Scott, M. Prockup, E. M. Schmidt, and Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," p. 12. II-C

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," pp. 556–562. III-A

[18] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," pp. 494–499. III-A, III-A

[19] P. López-Serrano, C. Dittmar, Y. Özer, and M. Müller, "Nmf toolbox: Music processing applications of nonnegative matrix factorization," 2019. III-A, III-A, III-D

[20] C. Dittmar, J. Driedger, M. Müller, and J. Paulus, "An experimental approach to generalized wiener filtering in music source separation," pp. 1743–1747, 2016. 5

[21] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984. 5

[22] O. Gillet and G. Richard, "Enst-drums: an extensive audio-visual database for drum signals processing," 2006. IV-A

[23] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep nmf for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70, IEEE, 2015. V