A STUDY OF INSTRUMENT-WISE ONSET DETECTION IN BEIJING OPERA PERCUSSION ENSEMBLES

Mi Tian[⋆]*, Ajay Srinivasamurthy*[†]*, Mark Sandler*[⋆]*, Xavier Serra*[†]

[⋆]School of Electrical Engineering and Computer Science, Queen Mary University of London, London
[†]Music Technology Group, Universitat Pompeu Fabra, Barcelona

m.tian@qmul.ac.uk, ajays.murthy@upf.edu, mark.sandler@qmul.ac.uk, xavier.serra@upf.edu

ABSTRACT

Note onset detection and instrument recognition are two of the most investigated tasks in Music Information Retrieval (MIR). Various detection methods have been proposed in previous research for western music, with less focus on other music cultures of the world. In this paper, we focus on onset detection for percussion instruments in Beijing Opera, a major genre of Chinese traditional music. A dataset of individual audio samples of four primary percussion instruments is used to obtain the spectral bases for each instrument. With these bases, we separate the input percussion ensemble recordings into its spectral sources and their activations using a Non-negative Matrix Factorization (NMF) based algorithm. A simple onset detection conducted on each NMF activation presents satisfactory overall detection rates, and provides us valuable implications and suggestions for future development of drum transcription and percussion pattern analysis in Beijing Opera.

Index Terms— Beijing Opera, Onset Detection, Drum Transcription, Non-negative matrix factorization

1. INTRODUCTION

Music Information Retrieval (MIR) is a recently emerging science focused on retrieving information from music using computer methods. Music is primarily an event-based phenomenon and detecting and characterizing musical events and their transitions is an important task in computer music applications. One intuitive strategy for content-based MIR research is to use musical concepts such as melody, rhythm or harmony to describe music.

Music transcription addresses the analysis of an acoustic musical signal so as to write down the pitch, onset time, duration, and source of each sound that occurs within it [1]. The automatic detection of onset events is an essential part of many music signal analysis algorithms and has various applications in identification, retrieval, musicological analysis, audio editing and coding, content-based processing and so on. The concept of a music onset in recent research is defined as the start of the transient, which marks the time interval during which the amplitude envelope of the signal increases. But this definition could become ambiguous in the case of the instruments having longer transient times without sharp bursts of energy rises. Vos and Rasch [2] approached this issue by introducing the concept of perceptual onset as the time when the most salient metrical feature of the music signal is perceived relative to its physical onset.

In the context of MIR, Bello et al. [3] reviewed some commonly used techniques for onset detection based on audio features such as the amplitude envelop, spectral magnitude and phase. Dixon [4] examined and proposed improvements to the then state of the art spectral methods. Klapuri [5] proposed a method utilizing band-wise processing and a psychoacoustic model of intensity coding to detect perceptual onsets. Though promising results have been achieved in percussion transcription [6, 7], state of the art music transcription and onset detection systems are still clearly inferior to skilled human annotation in their accuracy.

2. BACKGROUND

Most works on music transcription have focused on melodies of pitched instruments. However, recent years have witnessed a growing interest for transcribing non-pitched percussive instruments. The percussion instruments investigated in music transcription and onset detection tasks fall into two main types: membranophones, such as drums that have a stretched membrane or skin, and idiophones, such as cymbals that produce sound from their own bodies [8].

To address the problem of percussion transcription, some event-based systems [6, 9, 10, 11] have been proposed which segment the input signal into events informed by the percussion and then extract and classify features from these segments to uncover its musically meaningful content, such as onsets. An alternative to this approach is to rely on source separation-based methods to decompose the input audio signal into basis functions that capture the overall spectral characteristics of the sources. Commonly used source separation techniques and tools such as independent component analysis (ICA) and Non-negative Matrix Factorization (NMF) have proven to be useful in percussion onset detection tasks, especially when analyzing mixtures of different percussion instruments [7, 12, 13].

To the best of our knowledge, no work so far has attempted to study the transcription of Beijing Opera percussion, or the physics of its featured music from an MIR perspective. In this paper, we consider a Beijing Opera scenario and present an exploratory study on the use of NMF-based source separation techniques for instrument specific onset detection in Beijing Opera percussion ensembles, aiming at providing a baseline for further research in Beijing Opera percussion transcription.

2.1. Beijing Opera

We first provide a brief introduction to the music culture. Beijing Opera, also called Peking Opera, is a major branch of Chinese traditional music combining singing, dance and theatre art. Despite its

[†]This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). The authors thank Rafael Caro Repetto, MTG, UPF for his expert inputs on music concepts of Beijing Opera and Ying Wan from London Jing Kun Opera Association for his engaging performance during the recording of the dataset used in this paper.

rich musical heritage and the size of audience, little work has been done to analyze its music content from an MIR perspective. It has been included as a target in a few genre classification works [14] and the acoustical properties of Beijing Opera singing has been studied [15], but little work has been done to study the musical characteristics of its featured instruments.

The percussion ensemble in Beijing Opera establishes and maintains the rhythm in a performance and guides the progression of sections in an aria. Firstly, the percussion provides a base to indicate the rhythmic modes, called the *banshi*, and accompanies the singing voice. Secondly, the percussion ensemble plays different kinds of pre-defined, fixed, labeled patterns that create a context for different parts of the aria. These patterns are important indicators of the movements and actions of artists, the mood of the scene and cues to the scenario. An automatic description of these percussion patterns is thus quite important in providing the overall description of the aria. Percussion transcription is an important step towards describing these percussion patterns. For such a task, we first estimate the onsets of each of the percussion instruments, use this information in percussion transcription - which can be used for describing the meter and rhythmic progression of the piece, as well as describing percussion patterns that are important descriptors of the piece. It is with this goal that we explore instrument specific onset detection of Beijing Opera percussion and present this study.

There are six main kinds of percussion instruments in Beijing Opera, which can be grouped into four classes[1]. *Ban* (a wooden clapper) and *danpigu* (a wooden drum struck by two wooden sticks) are the primary instruments, referred together as *bangu* (clapper-drum). *Danao* and *qibo* (two cymbals that are collectively called "*naobo*" in general), *daluo* and *xiaoluo* (two gong instruments with different shape and pitch range) are the other main instruments [16, 17]. Since the sounds generated by ban and danpigu, and that of danao and qibo are very similar to each other and always happen at the same time, we follow the musicologically common practice to group ban and danpigu into a general class bangu, and group danao and qibo into a general class naobo. In our onset detection experiments, we will thus use these grouped classes as our target stimuli.

3. DATASET

Since there was no annotated collection of Beijing Opera percussion available, we built a dataset by recording sound samples with professional musicians in studio conditions at the Centre for Digital Music, Queen Mary University of London. The audio was recorded in mono using an AKG C414 microphone at a sampling rate of 44.1 KHz. The set of training examples used in the paper can be obtained from `http://compmusic.upf.edu/bo-perc-dataset`.

Dataset	Bangu	Daluo	Naobo	Xiaoluo	**Total**
Training	59	50	62	65	**236**
Test	1645	338	747	291	**3021**

Table 1: Beijing Opera percussion dataset showing the number of examples for each instrument in the training and test dataset.

The dataset, shown in Table 1, consists of recordings of the four percussion instrument classes: bangu, daluo, naobo and xiaoluo. Unlike pitched instruments, most idiophones cannot be tuned. These

[1] A few annotated audio examples of these instruments can be found at `http://compmusic.upf.edu/examples-percussion-bo`

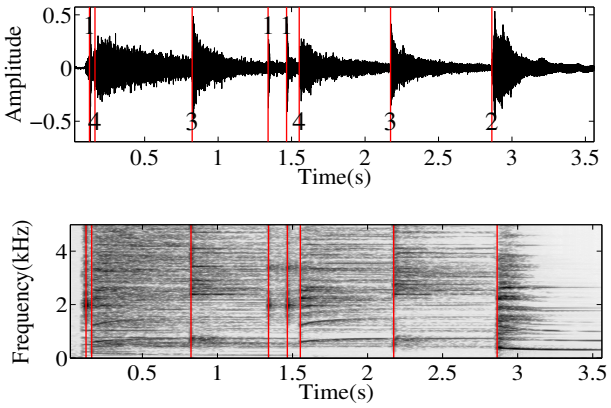


Fig. 1: An audio example containing all four instruments considered in the paper. The top panel shows the waveform and the bottom panel is the spectrogram, the x-axis for both panels is time (in seconds). The vertical lines (in red) mark the onsets of the instruments. The onsets are labeled to indicate the specific instrument onset: bangu-1, daluo-2, naobo-3, xiaoluo-4.

percussion instruments are made from metal casting or wood carving hence subtle differences might exist between the physical properties of individual instruments even of the same kind. For each kind of the above instruments, we recorded sound samples of 2-4 individual instruments played with different playing styles commonly used in Beijing Opera performances, hoping to achieve a better coverage of timbre and variations of playing techniques. The training set consists of short audio samples with single strokes of each individual instrument that capture most of the possible timbres of the instrument that exist in Beijing Opera. For the test dataset, we manually mixed the individually recorded instrument examples together using Audacity[2] into 30-second long tracks, with possibly simultaneous onsets to closely reproduce the real world conditions. The examples in training and test dataset are mutually exclusive.

Manual labeling of onset locations is tedious and time consuming, especially for complex ensemble music consisting of instruments with diverse properties. The onset ground truth was constructed by the taking the average onset locations marked by three participants without any Beijing Opera background. Participants were asked to mark the onset locations in each recording using the audio analysis tool Sonic Visualiser [18] displaying the waveform and corresponding spectrogram. Onsets generated by bangu have much lower amplitude and shorter transient time and happen in higher density than those generated by the cymbal instruments, and hence the bangu onsets are easily masked by the cymbals and gongs. Fig. 1 shows an audio example with all the four instruments. We can also see the amplitude dynamics and spectral shapes for each instrument. We can also see how the bangu stroke is masked by an adjoining xiaoluo stroke (0-0.5s in Fig. 1).

4. NMF-BASED SOURCE SEPARATION AND ONSET DETECTION

We describe the approach for instrument onset detection developed in our study. We use the training samples for each instrument and construct corresponding spectral bases using a NMF-based spectral

[2] `http://audacity.sourceforge.net`

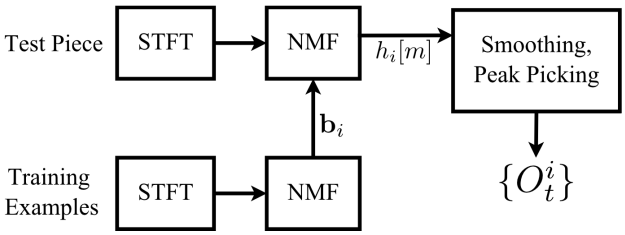


Fig. 2: Block Diagram of the Approach

decomposition. We then use these bases to separate the different instrument sources in a recording of the whole percussion ensemble, to obtain the activation function of each source. The activation function is then used to find the onset locations for each instrument. The process is outlined in Fig. 2 and is explained in detail in this section.

Non-negative Matrix Factorization aims to factorize a matrix $\mathbf{V}$ ($N \times M$) with non-negative elements into two non-negative matrices $\mathbf{W}$ ($N \times K$) and $\mathbf{H}$ ($K \times M$) such that $\mathbf{V} \approx \mathbf{W}\,\mathbf{H}$. For a spectrogram $\mathbf{V}$, the decomposition into matrices $\mathbf{W}$ and $\mathbf{H}$ can be interpreted as decomposing a spectrogram into a set of K basis vectors, grouped together in columns of $\mathbf{W}$ and a set of K activation functions corresponding to each of the basis functions in the rows of $\mathbf{H}$. Given a set of spectral bases, NMF provides a way to estimate the activation function for each of the basis functions over time. The activation functions thus obtained can be used to estimate events corresponding to each specific basis function. With this interpretation of NMF, we build a spectral basis for each instrument and use the activation function obtained using an NMF algorithm for onset detection. For an input spectrogram $\mathbf{V}$ and basis vector index $1 \leq i \leq K$, we represent spectral basis vectors as $\mathbf{b}_i$ such that $\mathbf{V} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_K]$ and the activation functions as $\mathbf{h}_i$ so that $\mathbf{H}^T = [\mathbf{h}_1^T, \mathbf{h}_2^T, \cdots, \mathbf{h}_K^T]$ where $(.)^T$ denotes the transpose of a matrix.

NMF algorithms are iterative and in this study, we use the original multiplicative update algorithm presented by Lee and Seung [19] that uses the normalized KL-divergence as its objective function. The implementation of the algorithm available in NMFlib v0.1.3[3] is used for both training and testing in our experiment.

The training audio samples described in Table 1 are used to estimate the spectral bases. We compute the spectrogram of each audio sample using a frame size of 1024 samples (23.2 ms) with a hop size of 256 samples (5.8 ms). The NMF of the spectrogram is used to generate a single spectral basis vector for the audio sample. The activation function is ignored. The basis vectors obtained from all the samples of each instrument are pooled and their mean is computed, leading to the basis vector $\mathbf{b}_i$ for the i^{th} instrument. We represent an instrument with a single spectral basis vector.

4.1. Decoding: Onset Detection

Given a test audio signal from a percussion ensemble, we first compute its spectrogram with the same parameters that we used for training. By using a fixed basis matrix $\mathbf{W}$ obtained from training, we use NMF to estimate the activation functions over time. With M frames in the spectrogram, we obtain the activation function for the i^{th} instrument as $\mathbf{h}_i = [h_i[0], h_i[1], \cdots, h_i[m], \cdots, h_i[M-1]]$, where m denotes a time (frame) index.

[3]NMFlib v0.1.3: `https://code.google.com/p/nmflib/`

The activation function is an important indicator of instrument onsets. Since we need to track major events in $h_i[m]$ using the peaks of the function, a smoothing is applied over the raw activation functions to aid peak picking. We use a gaussian smoothing window with a different standard deviation (σ) parameter for each instrument. Bangu sounds are sharp while the other sounds persist for a long time hence larger windows are needed. Based on these observations, in our experiments, we empirically fixed the σ to 5, 14, 11, 11 samples (which correspond to $\approx$ 29, 81, 64, 64 ms at the hop size of 256 samples) for bangu, daluo, naobo, and xiaoluo, respectively.

Peak picking on the smoothed activation functions provides us the onset time estimates for each instrument, during which we ignore low-amplitude peaks below 15% of the peak amplitude and ignore non-prominent peaks. A time difference threshold is also applied to select a single peak out of adjacent close peaks. We use a modified version of the peak estimation algorithm in VOICEBOX toolbox[4]. The set of peaks $\{O_t^i\}$ thus estimated from each smoothed activation function $h_i[m]$ for instrument i are used to evaluate the performance of the approach, providing insights into the nature of these percussion signals and suggesting broad directions to pursue transcription of Beijing Opera percussion ensembles more comprehensively.

5. RESULTS AND DISCUSSION

We report the accuracy of onset detection using the standard performance measures shown in Table 2. Given the possible inaccuracy of human annotation, a 50 ms time margin is given between the targets and detections in our evaluation. Onset detection for each instrument on an example ensemble excerpt from our test dataset is shown in Fig. 3. We observe a considerable improvement of peak picking performance both in amplitude and time after the smoothing is introduced. With smoothed activation functions, the peaks are better defined and important onset events are better estimated. For bangu which has the lowest energy in the ensemble, the smoothing function has made the onsets more salient and sharp. We also see that NMF performs well to separate the different instruments because of their distinct timbres.

Table 2 shows that a recall of 37.77% and an f-measure of 0.4495 is achieved on the whole test dataset. In general, the precision (55.50%) is higher than recall showing that many true onsets are missed. Of all the instruments, bangu has the best performance, with a high precision (71.69%). The performance with the two gongs daluo and xiaoluo is poorer with lower precision. Though precision for naobo (51.91%) is better than that for daluo (40.00%) and xiaoluo (31.49%), its recall (23.56%) is poor leading to an inferior f-measure (0.3241), hence the worst f-measure in the whole set of instruments.

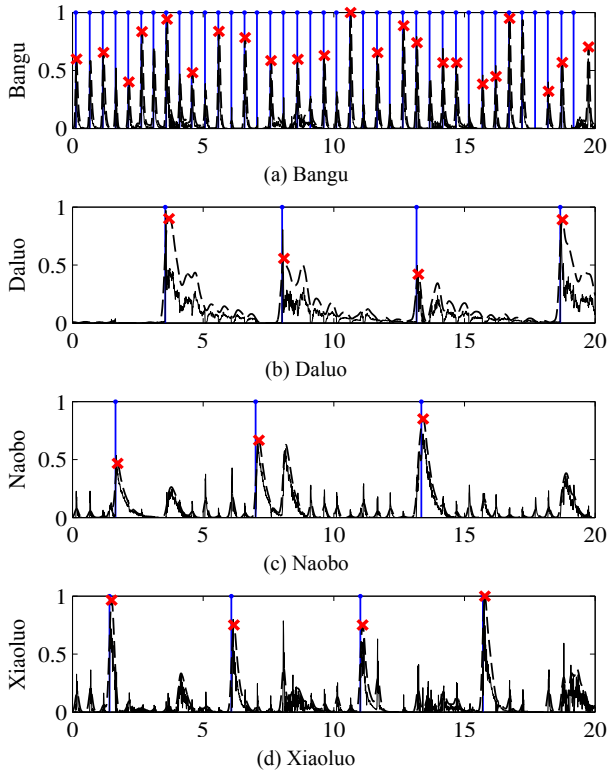
The precision being better in general shows that peak picking is effective in detecting strong onsets. However, the relatively poor recall rate is primarily due to volume dynamics and a non-adaptive peak picking algorithm. The changes in volume and amplitude through the piece can lead to varying peak amplitudes in the activation function. The low amplitude peaks are sometimes ignored by the peak picking algorithm (Fig. 3a). Significant improvement can be expected in recall rates if the peak finder adaptively chooses the threshold feeding with the local dynamics. This is an immediate logical extension to the algorithm that will be explored further.

The performance in general is better for bangu and poor for the gongs daluo and xiaoluo. Bangu has a sharp attack and a fast decay, and is short compared to the other instruments. Daluo and Xiaoluo

[4]`http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

Instrument	Targets	Correct	FP	FN	Precision (%)	Recall (%)	f-measure
Bangu	1645	699	276	946	71.69	42.49	0.5336
Daluo	338	152	228	186	40.00	44.97	0.4234
Naobo	747	176	163	571	51.91	23.56	0.3241
Xiaoluo	291	114	248	177	31.49	39.17	0.3491
Overall	3021	1141	915	1880	55.50	37.77	0.4495

Table 2: Results of onset detection for each instrument and overall performance on the test dataset. Targets: Total onset instances of the instrument in the test dataset, Correct: Number of correctly detected onsets, FP: Number of false positives, FN: Number of false negatives. Precision (P) and Recall (R) are presented in % values.



(a) Bangu (b) Daluo (c) Naobo (d) Xiaoluo

Fig. 3: Instrument-wise onset detection on a short excerpt from the test set showing the raw activation functions (black solid lines), smoothed activation functions (black dotted lines), estimated onsets ($\times$ in red) and the ground truth annotations (blue lines spanning the whole pane vertically), with the x-axis as time (in seconds).

have sustained sounds and a very characteristic time frequency signature (as can be seen from Fig. 1) that lasts over several time frames. We use a simple representation of the instrument basis in this paper, with a single time averaged spectral basis vector. The considerably poor performance implies that a single vector representation for daluo and xiaoluo is inadequate to capture the temporal evolution of the spectrum of the instrument sound and hence we need a better representation for the gongs. Since the gongs have a very characteristic time-frequency (TF) evolution, we can build TF templates for these gongs and use them for detection. This is another interesting direction to pursue - finding better representations for the different instruments under consideration. One such extension is to use convolutive NMF, which can work with complete spectral patch of many frames as a basis [12].

One important aspect we did not explore in this paper is the mutual dependence between activation functions. Activation functions are considered independently in the experiment, but each of the activation functions typically reveals information about the others. This can be explored further to perform a joint estimation of onset locations, with all the instruments together. A joint processing of the activation functions is a necessity for percussion pattern transcription since multiple instruments can play at the same time. Using local context information more effectively might also lead to a better performance instead of only tracking peaks. The peak shapes give abundant information about the type of onset, and hence modeling the peak shapes can be useful for improving the onset detection performance.

In summary, we get a better precision with the present approach but a poor recall. This can be mainly attributed to the dynamic volume changes and soft onsets that are not detected by the peak detector. The time-frequency characteristics of xiaoluo and daluo and the variety of timbral signatures of the naobo need to be explored further to develop better spectral representations to be employed by the NMF algorithms. Joint estimation of onsets utilizing context information and peak shapes are some other promising directions for our future work.

6. CONCLUSIONS AND FUTURE WORK

The final goal of our research is to describe percussion patterns in Beijing Opera. In this paper, we studied onset detection for specific instruments in the Beijing Opera percussion ensemble using spectral basis computed through an NMF-based approach. It is observed in our experiments that such an approach is promising, but further exploration is needed to present a concrete domain specific algorithm. We also proposed some broad directions for further exploration and to improve the performance. The time-frequency characteristics of xiaoluo and daluo and the variety of timbral signatures of the naobo need to be explored further to develop better spectral representations to be used in NMF algorithms, using multiple bases for each instrument. Adaptive smoothing based on peak shape modeling is also to be explored in the following study to improve the peak picking issues introduced due to the wide variety of onset types, variations in peak amplitude and contextual masking. Further, we aim to build a comprehensive dataset of samples of individual instruments as well as full length ensemble patterns that encompass different instrument timbres and playing styles. With such a dataset, we wish to explore further and develop specific algorithms for automatic percussion transcription and percussion pattern description in arias of Beijing Opera.

7. REFERENCES

[1] Anssi Klapuri and Manuel Davy, *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] Joos Vos and Rudolf Rasch, "The perceptual onset of musical tones," *Perception & Psychophysics*, vol. 29, no. 4, pp. 323–335, 1981.

[3] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 1035–1047, September 2005.

[4] Simon Dixon, "Onset Detection Revisited," in *Proc. of the 9th International Conference on Digital Audio Effects (DAFx'06)*, Montreal, Canada, September 2006, pp. 133–137.

[5] Anssi Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, USA, March 1999, vol. 6, pp. 3089–3092.

[6] Olivier Gillet and Gaël Richard, "Automatic transcription of drum loops," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, May 2004, vol. 4, pp. 269–272.

[7] Jouni Paulus and Tuomas Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proc. of the 13th European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, September 2005, pp. 4–8.

[8] Neville Horner Fletcher and Thomas Dean Rossing, *The Physics of Musical Instruments*, Springer, 1998.

[9] Fabien Gouyon, Perfecto Herrera, and Pedro Cano, "Pulse-dependent Analyses of Percussive Music," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, 2002.

[10] Masataka Goto and Yoichi Muraoka, "A Sound Source Separation System for Percussion Instruments," *Transactions of the Institute of Electronics, Information and Communication Engineers D-II*, vol. J77-D-II, no. 5, pp. 901–911, May 1994.

[11] Olivier Gillet and Gaël Richard, "Transcription and separation of drum signals from polyphonic music," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2008.

[12] Paris Smaragdis, "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," in *Independent Component Analysis and Blind Signal Separation*, Carlos G. Puntonet and Alberto Prieto, Eds., vol. 3195 of *Lecture Notes in Computer Science*, pp. 494–499. Springer Berlin Heidelberg, 2004.

[13] Samer A. Abdallah and Mark D. Plumbley, "Probability as metadata: event detection in music using ICA as a conditional density model," in *Proc. of the 4th International Symposium on Independent Component Analysis and Signal Separation (ICA 2003)*, Nara, Japan, April 2003, pp. 233–238.

[14] Yibin Zhang and Jie Zhou, "A study on content-based music classification," in *Proc. of the Seventh International Symposium on Signal Processing and Its Applications*, Paris, France, July 2003, vol. 2, pp. 113–116.

[15] Johan Sundberg, Lide Gu, Qiang Huang, and Ping Huang, "Acoustical study of classical Peking Opera singing," *Journal of Voice*, vol. 26, no. 2, pp. 137–143, March 2012.

[16] Yuan-Yuan Lee and Sin-Yan Shen, *Chinese Musical Instruments (Chinese Music Monograph Series)*, Chinese Music Society of North America Press, 1999.

[17] Elizabeth Wichmann, *Listening to Theatre: The Aural Dimension of Beijing Opera*, University of Hawaii Press, Honolulu, 1991.

[18] Chris Cannam, Christian Landone, Mark Sandler, and Juan Pablo Bello, "The sonic visualiser: A visualisation platform for semantic descriptors from musical signals," in *Proc. of 7th International Conference on Music Information Retrieval (ISMIR 2006)*, Victoria, Canada, October 2006.

[19] Daniel D. Lee and H. Sebastian Seung, "Algorithms for Non-negative Matrix Factorization," in *Proc. of Conference on Advances in Neural Information Processing Systems (NIPS)*, Denver, USA, 2000, vol. 13, pp. 556–562, MIT Press.