Separation of Tabla Strokes from Tabla Solo Audios Using Non-negative Matrix Factorization

A seminar report submitted towards partial fulfillment of the requirements for the degree of

Master of Technology

By

M A Rohit
Roll Number: 183076001

Under the supervision of:

Prof. Preeti Rao

Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India

May 2019

Abstract

This report concerns the problem of separating individual sounds of the tabla from a given sequence of overlapping tabla strokes, using non-negative matrix factorization based methods. The goal is to be able to provide a complete transcription of a tabla solo audio in terms of not only the onset times and labels of the strokes, but also additional information about the expressivity of the playing. Doing so would help understand the various ways in which strokes are played by artists to make the playing sound more artistic, and this knowledge could then aid a more natural synthesis of tabla solo audios. Source separation is needed because of overlapping sounds in the sequence resulting in inaccurate measurements on each stroke. And given the success of NMF-based methods in the source separation and transcription of the western drums, we examine its performance on the tabla.

Contents

1 Introduction

Sounds are a commonly found class of signals in our environment and society today. With speech and music being the most ubiquitous examples of sound signals that humans are constantly engaged in, a lot of research has gone into understanding the production and perception of these sounds. Although the more abstract nature and complexity of music gives rise to a different set of research challenges in music signal processing, common tasks in speech signal processing like automatic speech recognition, speech/speaker separation, speech synthesis, etc., have parallels in the field of music as well - viz., automatic music transcription, source separation in polyphonic music, music generation/synthesis, etc. The formal definitions of some of these terms appear below.

Automatic Music Transcription (AMT): AMT is the process of converting a music audio signal into a score-like representation that delineates the set of sound events occurring in it by specifying the onset time, duration, pitch and/or source of each one (Klapuri and Davy, 2007). A musical score (or a sheet) in the context of Western music, is a notation that is used by an artist to play/sing a composition. It contains essential details on how the composition must be played, like - the notes/strokes and their relative timings and duration, pauses, dynamics - how loud/soft a note should be, etc. However, automatic transcription systems till date have often only tried to produce the bare minimum transcription in the form of the onset times followed by labels that identify what the corresponding events are. As the authors in Benetos et al. (2013) note, much work is still needed to build systems that can provide a complete transcription consisting of information about dynamics, articulation type, and other indicators of expressivity that are commonly found in sheet music. Such a complete transcription system would be useful in musicological studies and help build more sophisticated automatic performance assessment systems. A big challenge faced in this regard however, is the lack of sufficient ground truth for such parameters.

Onset detection: Most musical events(sounds) have an energy envelope that follows a **attack** $\rightarrow$ **decay** or a **attack** $\rightarrow$ **decay** $\rightarrow$ **sustain** pattern, where the energy first increases to a maximum and then either drops quickly down to the noise floor or drops a little and then sustains for a short duration of time, while slowly fading away. This is illustrated in figure 1 with two different sounds. The plots above are the wave forms of the audio signals and the plots below show the time-envelopes of their energies with the different regions marked. The plots on the left(above and below) are of a sound with a sharp decay and negligible sustain, while those on the right are of a sound with a long sustain. The time period where the energy of the sound quickly rises and falls is termed the transient and is also defined as the interval where "the sound rapidly evolves in an unpredictable way" (Bello et al., 2005). The onset of the event is then an instant of time that marks some portion of this transient, and ideally coincides with the beat instants of the underlying rhythmic framework. It could be the instant when it begins, or when it reaches its peak. Accordingly, the onset could be computed as the time instant when the energy of the sound just crosses a predefined threshold in the rising phase, or is at its maximum.

Musical Source Separation (MSS): Musical source separation is the task of extracting the audio of any single source from a signal that contains different combinations of several sources playing simultaneously. An example would be - extracting the vocals from

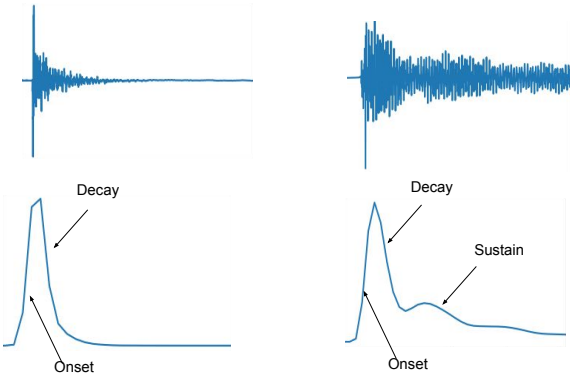


Figure 1: Different regions of a damped(left) and a resonant(right) sound

a recording of a song that contains drums, guitar, keyboard, etc. in the background. If the mixture contains as many channels as the number of sources and the mixing process is known and fixed, then the task may not be that difficult (Cano et al., 2019). However this is not the case in most practical scenarios, and MSS then involves understanding the unique properties and the structure of the individual sources and using this knowledge to break the mixture down into its constituent parts. MSS has some popular use cases, like generating karaoke tracks from recordings of songs, or extracting only the lead vocals to perform better pitch estimation for use in singing assessment tools. Another motivation for MSS is in cases like Indian classical music, where the underlying theory is scarce and a reasonable way to perform musicological studies is by analyzing recordings of performances. However an automatic analysis of these recordings is made difficult due to the presence of several sources - lead instrument/vocals, accompanying instruments, and the tanpura. Extracting the individual sources in such cases would not only help study each of them separately but also the relationships between them.

This report is about performing source separation on tabla solo audios, which are essentially sequences of different sounds produced on the tabla, as a means to producing a complete transcription. Although methods for tabla transcription achieving fairly good accuracies have been implemented, they do not focus on providing information beyond onset times and labels. The main aim of this work is to perform a more complete transcription by also recovering certain other properties of each stroke, e.g., its relative intensity, type of articulation, etc., so that the supra-segmentals - the expressive aspects of tabla music, can be better understood, as was attempted in Rohit and Rao (2018). This knowledge could then be used to perform more realistic synthesis of tabla sound sequences. Further, using a conventional method, like directly making measurements on stroke segments obtained from the audio, is prone to errors because of the following reason. The tabla contains a pair of drums each capable of producing sounds with a long sustain duration(resonant sounds). And since both the drums are not struck simultaneously all the time, it often so happens, due to the tempo of the playing being fast enough, that when one drum is struck before a previously produced resonant stroke on the other drum has completely faded away, these two sounds overlap. This makes accurate

measurement of parameters of individual strokes difficult, and one potential way to solve this is to separate the strokes before making the measurements so that the effect of the interference is reduced.

2 About the Tabla

2.1 The Instrument

Of the two drums that together form the instrument, the *bayan*, also known as the *dagga*, is bigger and produces lower frequency(bass) resonant sounds, and the *dayan*, also called the *tabla*, is smaller and produces higher frequency(treble) resonant sounds. Apart from the resonant sounds, both drums can also be struck in a number of ways to produce damped strokes - i.e., those with a sharp decay and no sustain. The characteristic nature of each drum stems from their unique construction. The *bayan* has a body made of aluminium, copper, steel or brass, and is completely hollow inside. The *dayan* on the other hand, is made of wood and may be completely or partially hollow inside. Both the drums have a membrane made of animal hide stretched on top with an additional annular layer of hide close to the circumference. Figure 2 shows the different regions on the membranes of both the drums.

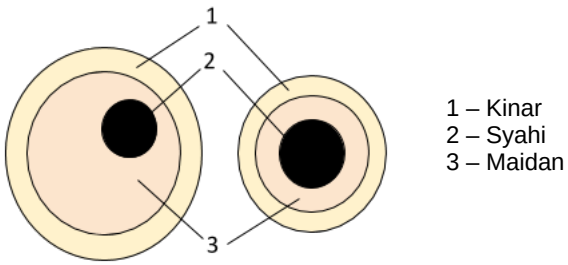


Figure 2: Regions of the *(*Bayan)(left) and the *Dayan*(right) membranes

A striking feature of these drums is the black patch on the membrane called the *syahi*, which is made from a mixture of iron filings, gum, soot, starch, and water. On the treble drum, the patch helps align the various partials produced by the membrane into harmonics and give rise to a sustained pitched sound. However, although the higher harmonics end up aligning quite well with tolerable deviation from the expected harmonic locations, the fundamental gets shifted up in the process by two semitones(i.e., to the second degree note). Hence, when the drum is struck such that the entire membrane can vibrate freely, by say, striking the membrane with the entire palm in an impulsive manner, two tones can be heard - one is the perceived fundamental because of the higher harmonics(the $2^{nd}, 3^{rd}, ..$) and the other is the actual fundamental that is two semitones higher than this. This can be seen in the plot of figure 3 that shows the spectrum of the sound produced when the membrane is allowed to vibrate freely by striking the membrane

and immediately lifting the palm after the strike. The spectrum shown is that of a short segment of the sound during the steady sustain phase after the onset. The peaks correspond to the fundamental and up to 4 harmonics above it. The dotted line is the F0 perceived due to the higher harmonics, while the tallest peak is the actual F0 that is 2 semitones higher than the perceived F0. On the *bayan*, the patch is placed slightly off the center and mainly only helps to reduce the fundamental frequency of the drum. The *dayan* is tuned to a musical note in the standard third or fourth octave, i.e., in the frequency range of 200-400 Hz, and the *bayan* usually has a fundamental around 100 Hz.

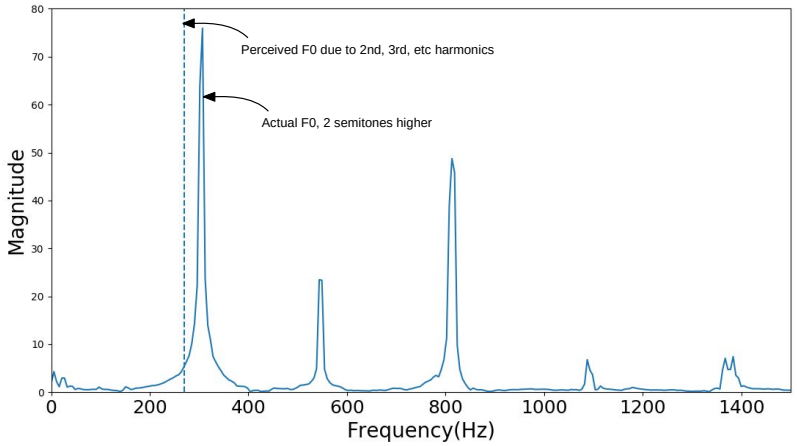


Figure 3: Magnitude spectrum of the sound produced when right drum is struck on its face with the entire palm in an impulsive manner, thus allowing the membrane to vibrate freely. The produced F0 is the highest peak(≈300Hz) and the perceived F0(≈270Hz) due to the rest of the harmonics is shown by the dotted line.

An important difference between the resonant sounds of the two drums is that the *dayan* can only produce sounds of a fixed F0, while the F0 of the *bayan* can be changed while playing a stroke. Although the tension of the membranes of both the drums can be adjusted so that their F0s are higher or lower, once the tuning is set, the resonant sounds on the *dayan* produce a fixed F0 that cannot be changed while playing a stroke. However, on the *bayan*, owing to its larger size and more loosely stretched membrane, the F0 of the resonant sounds can be varied by applying different amounts of pressure on the membrane using the wrist. This is called pitch modulation on the *bayan* or just *bayan*-modulation (Courtney, 2013).

The two drums can each be struck in a number of ways to produce distinct sounds, each with a unique timbre. These sounds can be broadly classified based on where they are struck - place of articulation(PoA), and how they are struck - manner of articulation(MoA) (Narang and Rao, 2017). A stroke may be played by striking any one of the three regions(shown in figure 2) alone or across more. The manner of articulation is usually one of two types - striking impulsively, ie., immediately raising the palm/finger(s) after the strike and thus allowing the membrane to continue vibrating and produce resonant sounds, or, striking and leaving the palm/finger(s) pressed on the membrane so that

the vibrations of the drum do not sustain, thus producing damped sounds.

Some sounds are produced by striking a single drum, others by striking both the drums simultaneously and a very few by striking both the drums but with a small delay between the two strikes. More interestingly, there exists a set of about 20 of these sounds (or strokes) that are recognised widely by tabla players everywhere and are referred to by onomatopoeic monosyllables that have no real meaning (Patel and Iversen, 2003). These syllables are called 'bols' and serve to identify the sounds. While it is remarkable that such a system developed in an age when classical music in India was not really formalised, it is perhaps not so surprising that it did, given how pedagogy in Indian music is predominantly an oral tradition. In any case, this practice of using bols makes it possible to give a representation to tabla music that is easy to communicate and remember. It has to however be pointed out that the mapping between the strokes and the bols is not always one-to-one. The same stroke may be called by a different bol and different strokes may be referred to by the same bol, all based on the context in which they occur. This is commonplace in tabla compositions but when the strokes/bols are considered in isolation, the mapping is fairly one-to-one. In the remainder of this document, the term bol and stroke may be used interchangeably.

Further, some strokes have a few variabilities in the hand gesture employed to produce the stroke, while maintaining the same PoA and MoA. Some of these variations perhaps arose due to different gharanas following different styles of fingering based on the kind of sound needed. However, not all of these variants sound equally distinct and some of them are alternate ways of playing to help play more easily in certain contexts. These variants are recognised as different ways of articulating the same stroke and are assigned the same bol.

Some details about the construction of the drums as described above are taken from Courtney (2016). A detailed description of each of the commonly found bols appears in the next section.

2.2 Tabla Bols

As mentioned earlier, each drum can produce sustained/resonant sounds as well as damped sounds, based on the way struck. This section introduces all the commonly found bols of each drum as well those that require the simultaneous playing of both.

2.2.1 The *Dayan*

The resonant bols on the *dayan* are - ***Na***, ***Tin*** and ***Tun***. ***Na***, also called ***Taa***, is one of the most commonly played bols. It is played by striking the *kinar* with the index finger while keeping the *syahi* muted with the ring finger to prevent the second degree note from sounding. This makes the sound get perceived only as the F0 of the higher harmonics. The strike is impulsive - the index finger is lifted immediately after the strike.

Tin is played like the *Na* by muting the fundamental, but by striking the *maidan*. It sounds similar to *Na* too, but has a slightly different distribution of energy across the harmonics, with the second harmonic being more energetic than the third, unlike in the *Na*.

Tun is sounded without muting any part of the membrane, and by striking the *maidan* to the right of the *syahi* such that the fundamental sounds more prominent than the higher harmonics that are also produced. A variant of *Tun* is ***Din***, which is produced by the striking the entire face of the membrane with the palm in an impulsive manner.

The damped strokes on the *dayan* are - ***Ti***, ***Ta/Ra***, ***Te***, ***Re***, ***Tak***, ***Tit*** and ***Da***. ***Ti*** and ***Ta/Ra*** are produced by striking the center of the drum and leaving the finger(s) pressed on it. The difference between the two is the particular finger used to strike the drum - the middle finger for ***Ti*** and the index finger for the ***Ta***. Two ways of playing *Ti* are commonly found - one using only the middle finger, and the other using the middle, ring and little fingers held together. The two variants have a subtle difference in their timbre.

Te and ***Re*** are also a pair like the *Ti* and *Ta/Ra* but they sound quite different and are also produced with different places of articulation and hand gesture. *Te* and *Re* are produced by striking the left half and right halves respectively of the drum, using the entire palm. Contrasted with the single finger used for *Ti* and *Ta/Ra*, *Te* and *Re* can certainly be expected to sound quite different.

Tak is also produced using the entire palm, but by holding it in a cupped shape and striking the *syahi*. This results in a very unique and sharp sound.

Tit employs a hand gesture similar to that of *Na*, but the finger is kept firmly pressed on the *kinar* after striking it, unlike in the case of *Na* where it is immediately lifted. It can thus be considered a muted version of *Na*. A variant of *Tit*, of the same name, is produced similarly but by striking the *maidan* and can be considered a muted form of *Tin*. Again, the two variants do not have a significant difference in their timbres.

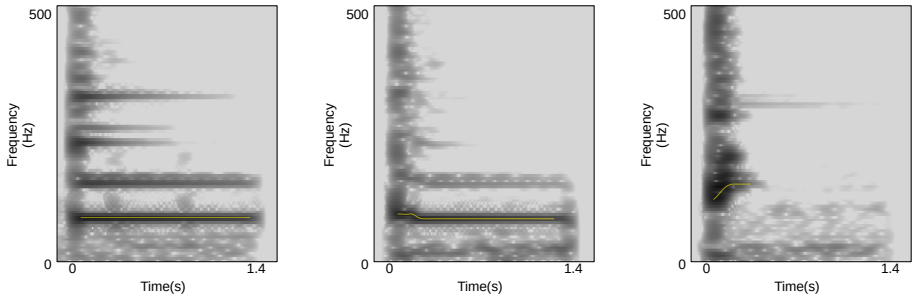
Da, despite being a damped stroke due to its manner of articulation, is not entirely devoid of a harmonic presence in its sound. This is because the stroke is produced by striking a small portion of the drum across the *kinar* and *maidan*, leaving a considerable portion of the drum undamped and free to vibrate, thus eliciting a faint harmonic sound. However, the sound has a sharp decay and sounds damped.

2.2.2 The *Bayan*

The two main bols of the *bayan* are ***Ge***, the resonant one, and ***Ke***, the damped one. Although the number of bols is fewer than the *dayan*, both of these bols have several variants.

Ge is produced by one of the following methods:

- Softly resting the wrist on the *syahi* and striking the *maidan* with either the middle and ring fingers together, or with the index finger alone. These two variants mainly exist to make faster playing easier for the cases when Ge has to be played several times consecutively in a sequence. They do not sound noticeably distinct from each other.
- Keeping the wrist firmly pressed on the *syahi* while striking, and then releasing the pressure immediately after the strike. This variant can also be produced using



(a) No modulation (b) Down modulation (c) Up modulation

Figure 4: Magnitude spectrograms of the three kinds of pitch modulation of *Ge*. (The faint line in white(yellow in the colored version) across the fundamental frequency shows its trajectory across time).

either the middle and ring fingers both or only the index finger.

- Striking the *maidan* with the wrist rested gently on the *syahi* during the strike, and sliding the wrist forward while firmly pressing down immediately after the strike. This is exactly opposite to the previous method. And again, as earlier, two variants exist based on the fingers used.
- Finally, by striking the drum with the palm close to the boundary of the membrane, on the *kinar*, thus allowing the entire membrane to vibrate freely and producing an extremely open sound. This variant is not used often while playing tabla compositions.

There is something more to be noted about the first three methods above. As mentioned briefly earlier, they correspond to different kinds of pitch modulation of the *bayan*. The first method produces a sound with a fairly static spectral character, ie., the fundamental frequency remains quite stable and the amplitude of the sound gradually decays. The second method produces a sharp fall (over time) in F0, due to the pressure being high just before the strike and low immediately after. The third method, on the other hand, exhibits a rise in F0 and the rate at which it rises can be controlled to an extent based on how quickly or slowly the wrist is slid forward. The three spectrogram plots in figure 4 illustrate these three methods, with each of them showing a small portion of the sound close to the onset, and up to 500 Hz on the vertical frequency axis. The fundamental of the drum on which these sounds were produced was about 80 Hz, as can be seen in the first plot, where all the harmonics remain at the same frequency throughout the sound. In the second plot, the F0 starts a little higher near 100 Hz and then falls quickly, immediately after the onset (the vertical region of uniform energy at the beginning). The third plot is quite different altogether with the F0 rising sharply during the onset to about 150 Hz, and then dying out immediately after that, and thus losing its resonant character.

Ke can be produced in one of the following three ways:

- Firmly striking across the *syahi* and *maidan* with the entire palm
- Firmly striking the *syahi* with only the top half of the palm (ie., all the fingers held together)

- Striking the *kinar* using the index finger in a snap like motion. This is done by holding it pressed to the thumb and then quickly moving the thumb away, thus releasing the index finger in a quick motion.

2.2.3 Compound (both drums)

The single drum strokes from both the drums are also played together as a combination sometimes. Although not every pair is commonly used, most strokes on the *dayan* are combined with the Ge (resonant stroke), and hence have at least one resonant component in the sound. These combinations are listed below:

- ***Dha*** - *Na* + *Ge*
- ***Dhin*** - *Tin* or *Tun* + *Ge*
- ***Dhi*** - *Ti* + *Ge*
- ***Dhit*** - *Tit* + *Ge*
- ***Tin*** - *Tin* or *Tun* + *Ke*

There are also combinations which consist of two strokes played in quick succession either on the same drum or one on each drum. By quick succession it is meant that at any given tempo of playing, they are more closely spaced than any two different strokes would be, and almost sound like they were struck together.

- ***Tra*** - played on the *dayan*, using strokes that resemble the *Ti* and *Ta*/*Ra*
- ***Kda*** - a combination of *Ke* on the and *Ta*/*Ra* on the *dayan*

2.3 Characteristics of Tabla Compositions and Solos

Tabla compositions are essentially sequences of strokes composed for the tabla and are filled with artistic and mathematical patterns. And like melodic compositions, they too are always set to a tala (a cyclic pattern of beats). While they can start on any beat of the cycle, they mostly end on the sam (the first beat of the cycle), often emphatically, signalling the end of the composition.

The scores for such compositions, apart from containing the sequence of bols in written notation, are often found with varying amounts of additional information. There is no established standard for how a score should be written, and artists and teachers employ their own methods as per convenience. However, if nothing else, there is usually some information in the score relating to the timing of the bols, e.g. the position in the cycle that the composition starts and ends at, the bols occurring on the same beat identified by writing them together, and separating such groups of bols using a punctuation mark(e.g., a long space) to denote beat boundaries.

Figure 5 below shows some examples of what typical scores look like. 5(a) was obtained from the booklet accompanying an educational video DVD of a tabla maestro orally reciting, playing and discussing several compositions from across different gharanas (Gupta et al., 2015). It has almost no additional information other than the bols, and it is not

clear whether the spaces between the symbols indicate beat boundaries or something else.

5(b) was obtained from Bel and Kippen (1992) and is more readable due to the consistent use of spaces to indicate beat boundaries. There is also a use of short forms for long phrases (Trkt in place of Ti-Ra-Ki-Ta) to make the writing appear more compact.

DHATITADHI TADHAGENA TINAKENA DHA-DHA-GENA DHINAGENA
DHATITADHI TADHA-GENA TINAKENA
TATITATI TATAKENA TINAKENA DHA-DHA-GENA DHINAGENA DHATITADHI
TADHA-GENA DHINAGENA

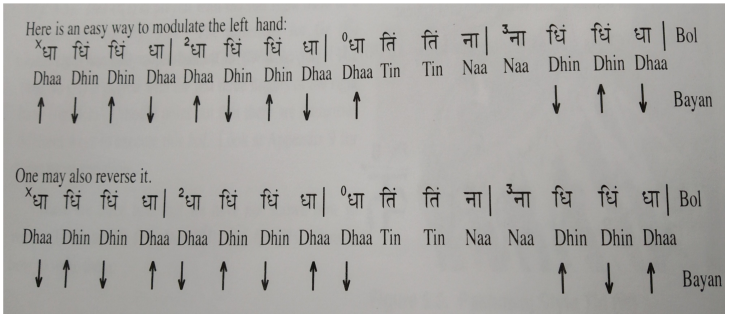
(a) From the accompanying booklet of a tabla solo DVD

dhatidhage	nadhatrkt	dhatidhage	dheenagena
dhatidhage	nadhatrkt	dhatidhage	teenakena
tatitake	natatrkt	tatitake	teenakena
dhatidhage	nadhatrkt	dhatidhage	dheenagena

(b) From a work on the structure of tabla compositions - Bel and Kippen (1992)

Dha Dha - Dha ; Dha Dha - Dhin ; Ghi,Da Na,Ga Ti,Ra Ki,Ta ; Taa Taa Ti,Ra Ki,Ta ;
Taa Taa - Taa ; Dha Dha - Dhin ; Ghi,Da Na,Ga Ti,Ra Ki,Ta ; Taa Taa Ti,Ra Ki,Ta ;
Dha Dha - Dha ; Dha Dha - Dhin ; Ghi,Da Na,Ga Ti,Ra Ki,Ta ; Taa Taa Ti,Ra Ki,Ta ;
Taa Taa - Dha ; Dha Dha - Dhin ; Ghi,Da Na,Ga Ti,Ra Ki,Ta ; Taa Taa Ti,Ra Ki,Ta ;

(c) From a previous work on expressivity in tabla playing - Rohit and Rao (2018)



(d) From a reference book on tabla - Courtney (2013)

Figure 5: Scores obtained from various sources

5(c) is from Rohit and Rao (2018) and is slightly more informative than the previous one, with the use of additional symbols like the hyphen, space and semicolon each with specific implications - the hyphen denotes a pause of the same duration as a single bol, a space just separates bols for improved readability and the semicolon indicates a beat boundary.

Finally, 5(d), obtained from Courtney (2016), seems to convey the most information. The 'x' mark, the numbers and the empty circles near the top left corner of some bols indicate the position in the cycle (the beat number) where that bol falls. More precisely, they mark the different sections of the taal (also called vibhaags). The small vertical lines just before the bols bearing these vibhaag markers also perform the same function - that of marking the vibhaag boundaries. And the spaces indicate beat boundaries. What is most interesting about this score is the presence of the upward and downward pointing arrow marks just below the bols. These pertain to the kind of modulation to be imparted to the *bayan* while playing the corresponding stroke. However, there are two sets of scores (a set being a line in Devanagari followed by English) in the figure and they are identical except for the directions of the arrow marks below the bols, ie., the kind of modulation imparted to the. The author notes that there is more than one valid form of modulation and one may follow either while playing that piece.

A tabla artist manages to impart a good deal of expressivity into the playing despite not explicitly putting any of this down in the score. An oral recitation of the composition is even more expressive and packed with prosodic variations. While the human voice is capable of imparting prosody in a number of ways, e.g., by raising or dropping the pitch or loudness, by changing the voice quality, etc., the tabla is quite limited in this respect. It has been observed in a previous work Rohit and Rao (2018) that there are two main indicators of expressivity in tabla playing - the intensity of the strike, and, if the resonant sound (Ge) is involved in the stroke, then the way it is modulated. The broader goals of the present work are to generate complete transcriptions of tabla audios that contain information on such expressivity as well and to attempt to derive a model for expressivity in tabla playing, if indeed there is one, by understanding the different ways in which strokes are played. And source separation is employed to ensure that the attributes of a stroke are measured accurately and are not affected by interference from a sustained previous stroke.

3 Non-negative Matrix Factorization

3.1 Introduction

Matrix factorization techniques are a popular class of algorithms used in machine learning problems to decompose and transform data into components that may have certain desirable properties. This process of decomposition is usually carried out by applying some constraints on the resulting components to make them more interpretable. Non-negative matrix factorization(NMF), for instance, requires that the components/factors have non-negative elements only. This offers certain useful representations in the case where the data is say, a magnitude spectrogram.

Consider an m x n matrix V of the magnitude spectrogram of an audio signal, where m is the size of the DFT and n is the length of the signal in number of frames. Then, NMF breaks V down into two matrices W and H, of sizes m x p and p x n, as shown in figure6 p is generally taken to be lesser than m and N, so that the decomposition results in a reduction of the original matrix.

This representation provides for a nice interpretation of the decomposition. Since W is

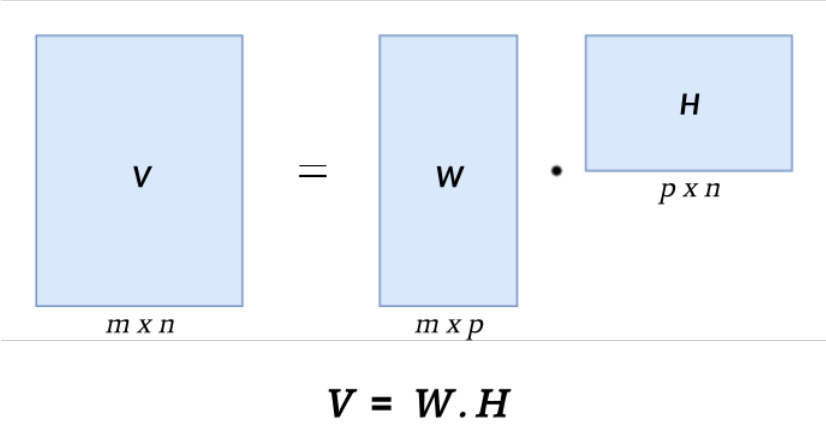


Figure 6: The basic definition of NMF shown pictorially

non-negative and has the same number of rows as V, its vertical axis can be taken to represent frequency, like in V, and hence each of column of W can then be viewed as a spectral slice, with the value in every element of the column representing the magnitude at the corresponding frequency. There are p such slices. And similarly, the horizontal axis of H can be taken to represent time(in frames), with the value in each element representing some kind of a positive scaling factor. And again, there are p such rows.

Each column W_i of W, (with $i = 1$ *to* P) gets multiplied with a corresponding row H_i in H and the resulting m x n matrix represents the contribution of this spectral slice across all the frames. Therefore, V can be viewed as a sum of p matrices of size m x n, each containing the contribution of one of p spectral slices across all the frames.

Another way to look at the decomposition is to consider the product of the entire matrix W with a column in H (Lee and Seung, 2001). The column represents the activations in a particular frame for every template. Therefore this multiplication operation is a linear combination of all the templates such that the entire activity in that frame/column of V is captured. The spectral slices are also called templates or bases and W is commonly referred to as the basis matrix; the rows of H are called activations (since they determine the activity of a template in any frame) and H is referred to as the activations matrix.

Therefore the goal is to find the optimum set of templates such that their linear combination using an appropriate set of activations explains the original data. The above is only the principle of NMF; the exact implementation details are described below.

3.2 Implementation of NMF

NMF is implemented as an iterative procedure where, starting with some initial values, W and H are iteratively updated such that the error between the the product $W.H$, and V, reduces with respect to some cost function. It is therefore common practice to represent the product $W.H$ as an estimate of V, and not as being exactly equal to it.

$$\hat{V} = W.H$$
$$V \approx \hat{V}$$

Some commonly used cost functions are:

1. Square of the Euclidean distance between V and WH:

$$||V - [WH]||^2 = \sum_{i,j}(V_{ij} - [WH]_{ij})^2$$

2. The Kulback-Liegler(KL) Divergence between V and WH:

$$D(V||[WH]) = \sum_{i,j}(V_{ij} \log \frac{V_{ij}}{[WH]_{ij}} - V_{ij} + [WH]_{ij})$$

3. The Itakura-Saito Divergence between V and WH:

$$D(V||[WH]) = \sum_{i,j}(\frac{V_{ij}}{[WH]_{ij}} - \log \frac{V_{ij}}{[WH]_{ij}} - 1)$$

Given one of these two cost functions, the objective then is to minimise the cost over all W and H subject to the constraint that $W, H \geq 0$.

As explained in Lee and Seung (2001), the above cost functions are convex in only one of the variables W or H and not in both together. It is therefore not easy to find a solution that finds the global minimum. However, numerical methods like gradient descent can successfully find the local minima. When working with such methods, speed of convergence and ease of implementation are important factors to consider. Hence, a set of multiplicative update rules that give a good compromise between these two factors are used instead of the usual additive updates found in gradient descent methods.

The update equations for all three cost functions can be expressed in terms of a parameter β, as given in Févotte and Idier (2011):

$$H \leftarrow H.\frac{W^T[(WH)^{\cdot(\beta-2)}.V]}{W^T[WH]^{\cdot(\beta-1)}} \qquad W \leftarrow W.\frac{[(WH)^{\cdot(\beta-2)}.V]H^T}{[WH]^{\cdot(\beta-1)}H^T}$$

where β is equal to 2, 1 and 0, respectively for the Euclidean distance, KL divergence and IS divergence based cost functions. The division is performed element-wise and so is every operation preceded by a '.' in the equations above.

In fact, there exists a more general class of cost functions specified in terms of this parameter β, where β can take any real value, called the beta-divergence. These cost functions reduce to the above three functions at the values 0, 1 and 2 (and upto a scale factor in the case of Euclidean distance). Although some research has taken place in experimenting with values of β other than the limiting values 0,1 or 2, it has been shown in Févotte and Idier (2011) that the cost function is not always convex for the other values, thus making them less favorable in tasks like musical source separation.

3.3 Non-negative Matrix Factor Deconvolution (NMFD)

The above formulation of factorization works well for sources whose sounds do not exhibit significant time-varying spectral characteristics. It has however been noted that a lot of musical instruments, including some drums, produce sounds that are not well represented by a single template. To cater to such cases, Non-negative Matrix Factor Deconvolution (NMFD) was introduced in Smaragdis (2004). This method also involves decomposing the data into templates and activations, but the templates are now time-frequency templates spanning more than one time frame and hence can represent time-varying content. In this method, it is assumed that the entire template occurs every time, as it is, without any variations. The decomposition is formulated as:

$$V \approx \sum_{t=0}^{T-1} (W_t \overset{t\rightarrow}{H})$$

where V, W and H are as earlier, but there are T number of W matrices, and the '$t \rightarrow$' on top of H indicates a shift operation of the columns to the right. An example of the shift operation appears below:

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \qquad \overset{1\rightarrow}{A} = \begin{bmatrix} 0 & a & b \\ 0 & d & e \end{bmatrix} \qquad \overset{2\rightarrow}{A} = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & d \end{bmatrix}$$

where a, b, c, d, e, f are some real numbers. The quantity T actually indicates the width(in number of frames) of every template and the matrix W_t is constructed by horizontally stacking the t^{th} columns of each template. The decomposition is therefore a sum of T matrices where each one contains the product of the t^{th} columns of all the templates and the same set of activations shifted to the right t times.

The update equations for a KL- divergence based cost function are the following (Smaragdis, 2004):

$$H \leftarrow H . \frac{W_t^\top \, [\frac{V}{WH}]}{W^\top \, 1_{m \times n}} \qquad\qquad W \leftarrow W . \frac{[\frac{V}{WH}] \, H^\top}{1_{m \times n} \, H^\top}$$

where $^\top$ is the matrix transpose operation and $1_{m \times n}$ is a $m \times n$ matrix of ones. Here too, the division is performed element-wise.

Although NMFD is useful to represent sources that have time-varying nature, it can also be quite restrictive in that a template has to always occur in its entirety. This means that if a sound occurs for variable durations of time, then it is difficult to represent it using a template of a fixed duration. This method therefore works best when a source produces sounds of roughly the same duration, regardless of the spectral variability over that duration.

3.4 Training

It was mentioned in the introduction of section 3 that the main objective in NMF is to find the optimum template vectors such that an optimum linear combination of them is "close" to the data matrix. Since the template vectors are much fewer in number than the number of data vectors, the hope is that they manage to capture some hidden structure

of the data (Lee and Seung, 2001) and hence generalise well (provided the data has such redundant structure in the first place). Further, since the algorithm iteratively converges to a local minimum, its performance can be enhanced if W and H are initialised appropriately. Hence, adding a training step before the decomposition, where templates are first learned from suitably chosen training data, could help decompose a test signal better. Furthermore, in cases where the templates are required to have a certain structure that needs to be preserved through the decomposition, implementing a training step may even be necessary.

The training data in such cases is generally chosen to be audios containing each source in isolation, from which one or more templates are learned for each source and saved in W. Then, given these "learned" templates, the decomposition could be implemented by modifying it to keep W constant and only update H. This method can work well if the templates are known to be reliably well learned and are required to retain that structure. If not, it could be performed as earlier, by updating both W and H.

3.5 Reconstruction

The task of source separation is usually encountered in scenarios where it is required to obtain separate audios for each of the multiple sources present in a mixture signal. In our case, we require separate audios of each of the basic strokes so that we can perform more accurate measurements on individual strokes. And since NMF only operates on the magnitude spectrogram, the decomposition only provides us the magnitude spectrograms of the separated audios. We need to somehow obtain the phase for each of these sources to be able to reconstruct their audios.

One way to perform the reconstruction is to retain the complex spectrogram if the input mixture signal and multiply it element-wise with a time-frequency mask for each source (Fitzgerald and Jaiswal, 2012). The idea then is to first generate such a mask for every source that attenuates the magnitudes of the frequency components that are not as significant for that source.

$$X_k = X \otimes M_k$$

where $\otimes$ denotes element-wise multiplication, X_k is the estimated complex spectrogram for source k, M_k is the time-frequency mask for this source and X is the original input complex spectrogram. In what is known as the generalised Weiner-filtering approach, the mask is derived as a ratio of the spectrogram of a source and the sum of spectrograms of all the sources.

$$M_k = \frac{S_k^r}{\sum_{n=1}^{N} S_n^r}$$

where S_k is the magnitude spectrogram of the k^{th} source (obtained from the decomposition), raised to some exponent r, and N is the number of sources, and division is performed element-wise. The exponent r is equal to 1 for magnitude spectra and 2 for power spectra. Due to the fact that the estimate obtained using the Weiner-filter mask is optimal in a least mean-square error sense, and given the success of divergence based cost functions in the decomposition step, similar divergence based methods have also been proposed to obtain the mask (Fitzgerald and Jaiswal, 2012):

$$M_k = \frac{D(S_k, WH)^t}{\sum_{n=1}^{N} D(S_n, WH)^t}$$

where S_k is the same as earlier, t, the exponent, is a tune-able parameter, D is a divergence measure, and again, division is performed element-wise.

An entirely different approach that does not make any use of the input spectrogram, relies on iterative estimation of the phase of a source given its estimated magnitude spectrogram, by using an algorithm like Griffin-Lim. However, since the input spectrogram is always available in musical source separation tasks, this method has been found to be used rarely.

4 Review of NMF-based Methods in Drum Transcription and Separation

The vast literature on music signal processing is replete with research on several western musical instruments, but not so much on Indian instruments. While the tabla has been studied to understand its working and the physics of its sound production (Raman, 1934; Ramakrishna and Sondhi, 1954), performance aspects of solo tabla playing are yet to be studied thoroughly. And when it comes to tabla transcription, the most researched methods seem to be built on principles often found in speech processing - using certain well-chosen features and a trained classifier to identify each entity, using a temporal language model to improve the acoustic model, etc. While these methods have proven effective for a transcription task, they have not been used to provide a complete transcription as was described earlier. These methods may even be insufficient for such a task due to the interference between adjacent strokes, thus requiring a separation of the strokes before further measurements can be made. Hence, we turn to the case of the western drums, which, despite being quite different from the tabla, have enough similarities with it, and more importantly, have been subject to the application of NMF for separating the different drums from polyphonic[1] audio. However, it has to be noted that most of the research on drums has used audios containing 3 drums - the kick/bass drum, the snare drum and the hi-hat, with an objective to separate these three drums. In this regard, the problem to be solved in the case of tabla is different in that we wish to separate not just the two drums(*bayan* and *dayan*) but also the timbrally different strokes of each drum.

Section 4.1 describes the various works in literature that have used NMF for automatic drum transcription and drum separation tasks. Studying this literature can provide useful insights about the separation algorithms that can be used for the tabla and the challenges they pose. Section 4.2 reviews some of the past work on tabla transcription.

4.1 Transcription and Separation of Western Drums

An early work that used NMF for drum transcription is Paulus and Virtanen (2005). Here, a training step was included prior to the decomposition of the test sequence, in which isolated instances of every drum were used to obtain template spectra that were then kept fixed during test audio decomposition. The training data contained 20 isolated instances of each of three drums - snare and bass drums and the hi-hat. The prior template for

[1] Polyphonic - multiple sources producing sound simultaneously

every drum was obtained by applying NMF on each instance and then averaging across the 20 instances. The test data was a set of recorded drum sequences containing all three drums. Upon decomposing a test sequence, the resulting activations were used for onset detection after some post-processing. The best transcription performance was obtained using the KL-divergence based cost function. The following interesting observations were made:

- working with a coarse frequency resolution of the STFT was better as drum sounds tend to have some stochastic nature that becomes more apparent at finer frequency resolution
- the method did not work as well on more complex signals containing melodic[2] instruments along with drums, since the data in such audios did not fit the model well

A related work from the same group was on separating drum sounds from a recording containing drums and melodic instruments (including vocals) (Helen and Virtanen, 2005). Here, a test signal containing a mixture of harmonic and drum sounds was decomposed into a set of 20 templates and corresponding activations. Each of the templates were then classified as belonging to either a harmonic class or drum class by using an SVM classifier. The classification was performed using a set of spectral features derived from the templates as well as some temporal features from the activations. The SVM was pre-trained on several separate instances of harmonic and drum sounds. Finally, using the separated magnitude spectrograms from the decomposition and using the phase from the input signal, separate audios were reconstructed. Although this work only attempted to identify a given sound as belonging to a drum or a harmonic source, and did not try to identify the particular kind of drum that produced it, it is interesting because if enough training data were available for each of the individual drums, then this method could be extended to accomplish that as well. However, the availability of such data can be a big challenge, especially in the case of tabla.

Offering a deviation from the usual single channel settings was Alves et al. (2009), which attempted to perform NMF based decomposition on multichannel drum recordings, where each drum was recorded using a separate microphone. The extension from the single channel settings was to use a different template for each drum in every channel, such that, if there were N channels and C sources, then the total number of templates was $N \times C$. These templates were stacked to form a single matrix and used to decompose the matrix which had the signal spectrograms from each channel also stacked. It was found that using multiple channels did improve the performance over using a single channel.

A more recent work addressing separation and transcription of drums from drum solo recordings is Dittmar and Gärtner (2014), where the initial training step obtained prior templates for each drum by simply averaging across the spectrograms of its isolated instances. Then, three ways of performing NMF decomposition of test audios were implemented and compared:

1. The prior templates were kept fixed and only the activations were learned, as in the earlier works.

[2]Melodic - instruments that produce pitched, harmonic sounds

2. The template matrix W was allowed to freely get updated, with the prior templates only acting as a good initialisation.

3. A combination of the previous two, called "semi-adaptive" in which, the template matrix W in every iteration was obtained using a combination of the prior and and the update, ie.,

$$W = \alpha.W_p + (1 - \alpha).W$$

where W was obtained from the update equation of the current iteration and W_p was the matrix of prior templates. α, called the blending parameter was defined as below:

$$\alpha = (1 - \frac{k}{K})^{\beta}$$

where k was the current iteration number, K was the iteration limit (total number of iterations) and β was a tunable parameter. This form for the update of W ensured that the estimated templates were closer to the prior templates at the start of the decomposition and slowly adapted to the test data as the iterations progressed. This allowed the additional variations in the test audio that had not got captured in the prior templates to get learned and hence improve the decomposition. And it was found that this method indeed worked better than the other two.

An interesting work that introduced variations to NMF in order to overcome the lack of temporal structure in the templates was Battenberg et al. (2012). Here, drum sounds were separated into "head" and "tail" regions corresponding to their onset and decay portions, and separate templates were obtained for each region. An electronic drum kit was used to record data and the training set contained several instances of each drum played at "a variety of dynamics and articulations" (Battenberg et al., 2012). For each drum, different sets of head and tail templates were obtained by clustering all the spectral slices of the corresponding regions and using the mean vector from each cluster as the representative template. The number of templates for each drum was not constrained to be the same and was chosen based on the number of clusters returned by the clustering procedure. Finally, during the decomposition of a test audio, the templates were kept fixed and only activations were solved for. Since the data contained strokes played at different dynamics and was generated using an electronic drum kit, the ground truth intensity levels were also available. The separation quality was thus evaluated by comparing the obtained activations with this ground truth. A similar work is Downing, an unpublished report) where the templates were simply a time-average of the spectrogram across all the frames containing only the head or tail regions. And then, during test audio decomposition, these templates were allowed to adapt in a manner similar to the "semi-adaptive" method described earlier.

NMFD has also been used abundantly in drum transcription and separation tasks. In fact, the work that formulated NMFD (Smaragdis, 2004) also tested it on drum audios. But it was quite preliminary as the audios did not contain significantly overlapping sounds. Besides, the aim of the work seems to have been to primarily evaluate the templates learned in terms of how well they represented the individual drums. This was also verified by subjectively evaluating the quality of the reconstructed audios of each of the drums after

the separation.

In Lindsay-Smith et al. (2012), the authors explored the use of NMFD for drum transcription and offered a modification to the update expression for the activations H to make them more impulsive. They also compared the use of STFT against a mel spectrogram as the input representation and found that the former yielded better results. However, it was pointed out that the better performance with STFT was at the cost of increased time for the decomposition, which can be an issue for real-time applications. Training and testing data was generated using a software with high quality drum preset sounds. Testing was performed on three sets of data - one with simple loops containing only the three drums - kick, snare and hi-hat, another with more complex loops containing two more drums - cymbals and tom-toms, and yet another with even more complex loops containing the five drums along with two different kinds of articulation on the snare and hi-hat. The performance was found to be the poorest on the last set.

A recent work using NMFD for source separation is Dittmar and Müller (2016), where some new techniques were proposed to guide the decomposition and hence improve the performance. One of these was similar to the method used in other works above - providing prior templates. The other improvements were based on the assumption that the score or ground truth transcription for a test audio was available. Using this, the ground truth onset locations were used to initialise the H matrix, by first planting an impulse at each of these locations in H, and then using an exponential moving average filter to impart a slow decay to each of these impulses. And due to the use of multiplicative updates, entries in H that were initialised to zero remained so after the decomposition as well. This was found to minimize the effect of crosstalk when given a good set of prior templates. Further, the decomposition was performed in two stages, with the first one as usual - using well initialised W and H and obtaining separated audios after decomposition. Then, the fully updated W was kept fixed and reused to separate the individual separated audios again, in order to further eliminate cross-talk. Another significant contribution of this paper was the use of objective metrics to evaluate separation quality. And in particular, metrics that evaluate the quality of the individual reconstructed components based on the level of cross-talk were favored.

4.2 Review of Previous Tabla Transcription Systems

The first work on tabla transcription was Gillet and Richard (2003), where the system consisted of an automatic onset detection step to segment the test audio into individual bols, followed by a tempo detection step to find the underlying grid and correct errors in the onset detection, then a feature extraction step and finally a classification step using classifiers like k-nearest neighbors and a naive bayes classifier. A hidden markov model was also included to make use of temporal dependencies and boost the performance. Chordia (2005) was an extension of this work where the database was expanded and a few other classifiers - neural networks and decision trees were experimented with, reaching higher accuracy levels. While these accuracy levels were good compared to the state of the art in other music transcription systems, it was found in Chordia and Rae (2008) that they were not good enough for real time applications. It was also noted that the most common error in transcription was in the case when a damped stroke on the *dayan* was preceded by the resonant stroke *Ge*. This is in line with our observations and further motivates the use of source separation techniques.

5 Dataset

5.1 Recorded

A dataset of solo tabla compositions was prepared by getting a tabla artist to orally recite and play them. Recordings were carried out in an anechoic chamber using an electret measurement microphone. The microphone was placed about two feet away from the tabla at roughly the same height as the instrument. The artist listened to a lehra(the equivalent of a metronome) on headphones as the reference tempo to recite and play on. Each composition was first recited and then, after a pause of one cycle, played on the tabla. There are 25 compositions and they are all set to the sixteen-beat cycle tintal. They are mostly kaidas, with a few chakradhars and gats. Kaidas are extendable compositions with a theme and variations pattern with the constraint that the variations must only use bols from the theme (Pradhan, 2011). In practice, the theme, which usually spans one or two cycles and is the identifier of the kaida, is played first and is followed by several variations. For the present dataset, only the theme of every kaida was recorded. Chakradhars and gats are non-extendable pieces composed by great maestros of tabla music and these are between 2 and 8 cycles long in the dataset.

The audios of the recitation and the tabla playing were then annotated as follows. Short-term magnitude spectra were computed across each audio over windows of size 20 ms and a hop-size of 10 ms. These were differenced across adjacent frames to be used as a measure of 'spectral flux', as is commonly done in musical note onset detection (Bello et al., 2005). We applied the same to our recordings and used peaks in the spectral flux to mark the syllable and stroke onsets. To avoid false alarms, adjacent peaks closer than 40 ms were combined into a single onset. Then, the set of onsets of each composition were time-aligned with the corresponding bol sequence obtained from the written composition to achieve the automatic segmentation and labeling of both the recitation and tabla playing recordings.

Further, a set of isolated strokes played on the same tabla-set was also recorded, for which the artist played each of the distinct bols described in section 2.2, in isolation, such that there was no overlap between consecutively played sounds. This was ensured by waiting till a sound had completely faded away and then playing the next bol. As described earlier, some bols can be played in multiple ways, either with a different hand gesture while keeping the PoA and MoA the same, or with a similar gesture but by striking a different part of the drum. All such variations that are commonly played were also recorded, making the dataset quite complete in terms of the inventory that is required to play common tabla compositions. Furthermore, to capture the variations related to expressivity viz., intensity and *bayan* modulation as explained in section 2, every stroke was played at two levels of intensity, one low and the other high, and strokes that involved the use of *Ge* were played with different kinds of modulation of *Ge*. Finally, to capture any uncontrolled or unpredictable variations in stroke articulation that may occur in practice, every stroke was played thrice (at each intensity level and each variation). The table below provides a summary of the dataset.

Bol	Drum	No of types	No of utterances		Nature of variations
			Soft	Loud	
Na	*Dayan*	1	2	3	
Tin			4	4	
Tun			4	3	
Din			3	4	
Ti		2	6	7	Different fingers used
Ta		1	4	5	
Te			3	2	
Re			3	2	
Da			6	4	
Tak			3	3	
Tit		2	6	8	Different places of articulation(PoA)
Tra		1	3	4	
Ge	*Bayan*	5	20	10	Different fingers used, and played with and without pitch modulation
Ke		3	12	13	Different PoA and hand gestures used
Dha	Both	2	6	5	With and without Ghe-modulation
Dhin - Tin+Ge			6	7	With and without Ghe-modulation
Dhin - Tun+Ge			6	7	With and without Ghe-modulation
Dhi		1	3	3	
Dhit		4	11	-	Variations found in Tit + variations in Ghe-modulation
Kda		1	3	5	

Table 1: Table showing the number and kind of variations of each stroke in the dataset of isolated strokes

5.2 Synthesized

Equipped with the database of isolated tabla strokes, each of the compositions that were played by the artist were then synthesized, in order to create a validation set to tune our algorithms on. Working with synthesized compositions gives us the chance to define a ground truth which is otherwise hard to generate for the recorded audios. By precisely controlling the kind of variability present in every stroke we can understand the performance of the algorithms on a case-by-case basis. However, it has to be noted that although the artist recognised the database as being exhaustive in terms of both the bols and their variations, it is possible that while playing the compositions, more variations are introduced due to some influence of adjacent strokes. It is also possible that every stroke might not always be played exactly at one of the two levels of intensity that were recorded for the dataset, and may not also just be a scaled version of one of these. In such a case, the observed stroke in the playing may be some combination of the two. Therefore, despite being extremely useful, the database by itself may not yet be as complete as needed to handle compositions played by an artist.

In the process of synthesis, the first step was to generate a transcription file containing

the exact onset times and the labels of the corresponding bols. This was generated in two ways. One, by parsing the score of the composition to get the sequence of bols and information about the number of bols occuring on every beat of the cycle, and then creating a transcription file using an arbitrary tempo close to that of the played realisation (this method offers the freedom to change the tempo and thus generate more data for analysis). And two, by simply generating a transcription of the recorded audio itself.

The next step was the synthesis, where, every stroke was picked from the database and placed in an audio stream such that the onset of the isolated stroke aligned with the onset time in the transcription file. While doing so, care had to be taken to choose an appropriate duration of the isolated stroke. This was not an issue in the case of damped strokes because they do not last long anyway, but the resonant strokes in a real scenario end as soon as the drum that produced them is struck again to play some other stroke. Therefore, in the synthesis, after a resonant stroke, say 'A', was produced, every new stroke 'B' had to be checked to find out if it was from the same drum as A and if so, then A was ended where the onset of B was. Else, if B was produced on the other drum, then A was allowed to be present in the stream, along with B.

There was still one hurdle which could not be overcome. When a compound resonant stroke, like say, *Dha*, which is made of two resonant strokes *Na* and *Ge*, is followed by any stroke on only one of the drums say, the *dayan*, then in a real scenario, only one of the resonant strokes in *Dha*, in this case only the *Na*, would stop and the other, in this case *Ge*, would continue into the next stroke. Although our database of isolated strokes has separate instances of all the individual strokes that make up compound strokes, it is not yet known how a compound stroke can be accurately synthesized from them. Therefore, during synthesis, recordings of compound strokes (also present in the database) were used, and when a scenario as described above occurred, we were unable to stop only one of the resonant strokes. This can be fixed by understanding how a compound resonant stroke can be synthesized using its constituents.

6 Experiments

Motivated by the success of matrix factorization techniques for automatic drum transcription and separation, NMF algorithms were implemented for separation of tabla strokes. The database of isolated strokes of the tabla was used to learn templates and different training methods were experimented with. The number of sources was set to be equal to the number of distinct single drum strokes, and the compound strokes were expected to be learned as a combination of the constituent individual strokes. Finally, reconstruction of separated audios was achieved using the Weiner filtering method. Figure 7 shows a block diagram of the whole process, inspired by the system in Dittmar and Gärtner (2014), with the following steps: learning the templates, using them to decompose a given composition audio (with or without the help of the score) into the basic strokes, using the resulting activations and templates to reconstruct separated audios for each of these strokes and finally performing transcription and making intensity and modulation measurements using either the separated audios or the output activations. Since NMF has not been applied earlier to tabla solo audios in this context, it is yet to be seen whether the output activations or the reconstructed audios yield better performance in the final

tasks (hence the final stage in the block diagram shows a switch).

The next few sections describe the various methods implemented for training and decomposition, based closely on methods from the literature on drum transcription and separation.

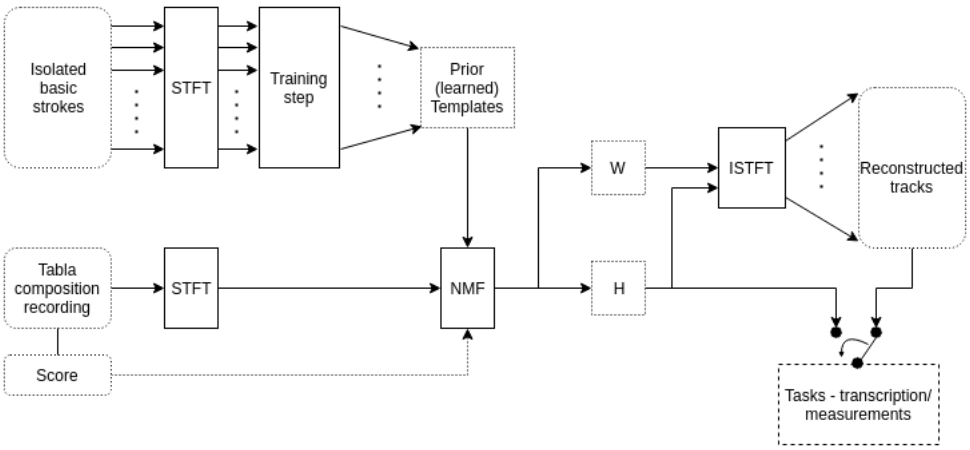


Figure 7: The entire pipeline - a training step to learn prior templates, using them to decompose a test audio, possibly along with the score of the composition, and then using either the output activations or the reconstructed audios for further measurements/transcription

6.1 Training Step

The following training methods, based closely on the methods in the literature reviewed earlier, were implemented:

1. All the isolated instances of a stroke were concatenated into a single audio signal, on which NMF decomposition was then performed to learn multiple templates per stroke.

2. Based on the use of head and tail regions in drum sounds, onset and decay regions were detected automatically for every stroke using empirical thresholds on short time energy and a method similar to point 1 was implemented to learn two templates for every stroke separately from these two regions. The templates were obtained as the average spectral slice across the corresponding region and then averaged across all the instances.

6.2 Decomposition Step

NMF methods were implemented as outlined in section 3 and each of the following modifications were compared:

1. With and without the use of available score(ground truth transcription). The score was used to determine whether a frame contained a previous resonant stroke that had not yet ended. Then, the decomposition was performed at the stroke level using all the templates of only the current stroke and the particular sustained resonant stroke present at the same time.

2. When the templates were separately learned for the onset and decay regions, the decomposition was again performed at the stroke-level using all the templates of the current stroke but only the decay templates of any previous stroke that was also present during the current stroke.

7 Evaluation: A Case Study of 'Dhin Na Ge Na'

A qualitative evaluation was performed to compare the various methods by testing the quality of separation in a small phrase of 4 bols: '*Dhin Na Ge Na*'. This phrase is representative of the case of overlapping strokes and is also quite commonly found in several compositions, thus making several instances of it available. All the strokes in this phrase are resonant, with *Dhin* being a compound stroke containing the individual strokes *Tun* and *Na*. The spectrogram of a one second long instance of this phrase taken from one of the recorded audios appears in figure 8, with the note onsets marked right below it. the It can be seen that the *Tun* (tone at around 300Hz) in the *Dhin* ends as soon as the next *Na* is played, while the *Ge* (tone at around 100Hz) sustains until the next *Ge* is produced. And during the (second) *Ge*, the previous *Na* sustains, as seen by the presence of harmonics between 500 Hz and 1500 Hz. These harmonics are obscure during the onset of this second *Ge* but quite prominent during its sustain. And then finally, *Ge* sustains while the *Na* stops and is replaced by another instance of *Na*.

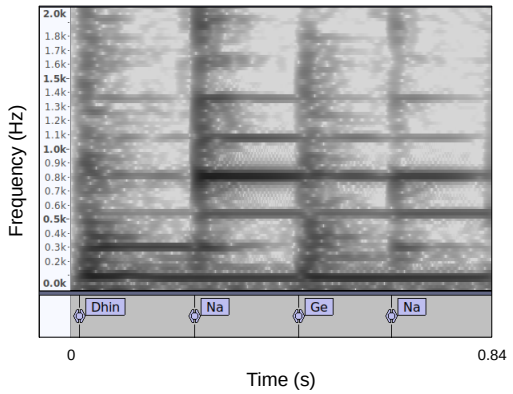
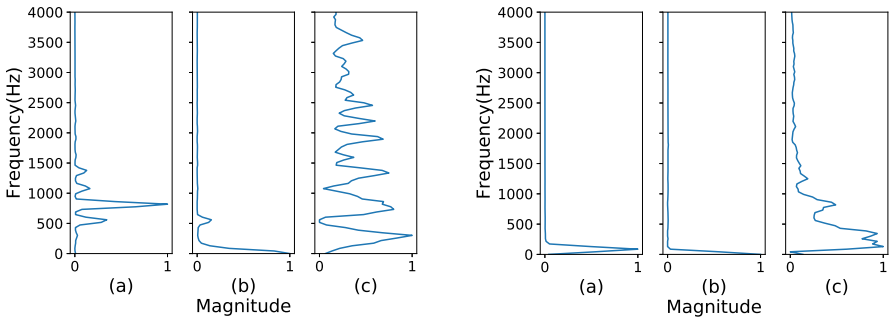


Figure 8: Magnitude spectrogram of an instance of the phrase '*Dhin Na Ge Na*'

The various separation methods described previously were evaluated qualitatively in the following ways:

1. By analysing the spectral content of the learned templates(in the training step) and how well they matched with that of the corresponding strokes



(a) Templates for *Na*

(b) Templates for *Ge*

Figure 9: Prior templates learned in the training step for two strokes - *Na* (resonant stroke on the *dayan*) and *Ge* (resonant stroke on the. The columns (a)-(c) in each plot are the three templates learned.

2. By comparing the spectrogram plots of the separated audios of the phrase '*Dhin Na Ge Na*' by decomposing using different methods

7.1 Evaluating the Trained Templates

Figure 9a shows three templates learned for the strokes *Na* and *Ge*, in the columns (a)-(c), by decomposing sequences of several isolated instances of each of them. Each template has been shown in the range of 0-4kHz range. The first template in each case seems to have picked up the harmonic components of the sounds, with the peak frequencies matching those observed in the actual spectra of the sounds. The third template of *Na*, despite having some prominent peaks, has a noisy broadband structure that most likely corresponds to the onset region of the stroke. The same in the case of *Ge* has a slightly wide band at a low frequency perhaps owing to the continuous frequency sweep resulting from the modulation. The middle template does not contain much information in either case. Thus, increasing the number of templates could help break these templates down further into more elementary forms and allow better decomposition. But keeping the number too high might make the templates less interpretable.
Similar templates were obtained by learning templates separately on the onset and decay regions. The advantage with this method was that the template belonging to each region was known beforehand and therefore decomposition of every stroke could be performed by controlling the particular templates used.

7.2 Evaluating the Separation Quality

Figure 10 shows the spectrograms of the audios obtained for the individual bols after decomposing the phrase using three methods. The phrase contains only three distinct bols - *Na*, *Ge* and *Tun* and these appear at the top of the plots. Further, to observe the extent of cross-talk, three other strokes are also shown: *Tin* - a stroke whose timbre closely resembles that of *Na*, *Ke* - a damped bol on the *bayan*, and *Ta* - a damped stroke on the *dayan*. Each vertical set of the plots corresponding to each of these strokes was obtained from a different method: in (i) using several templates per stroke but without using

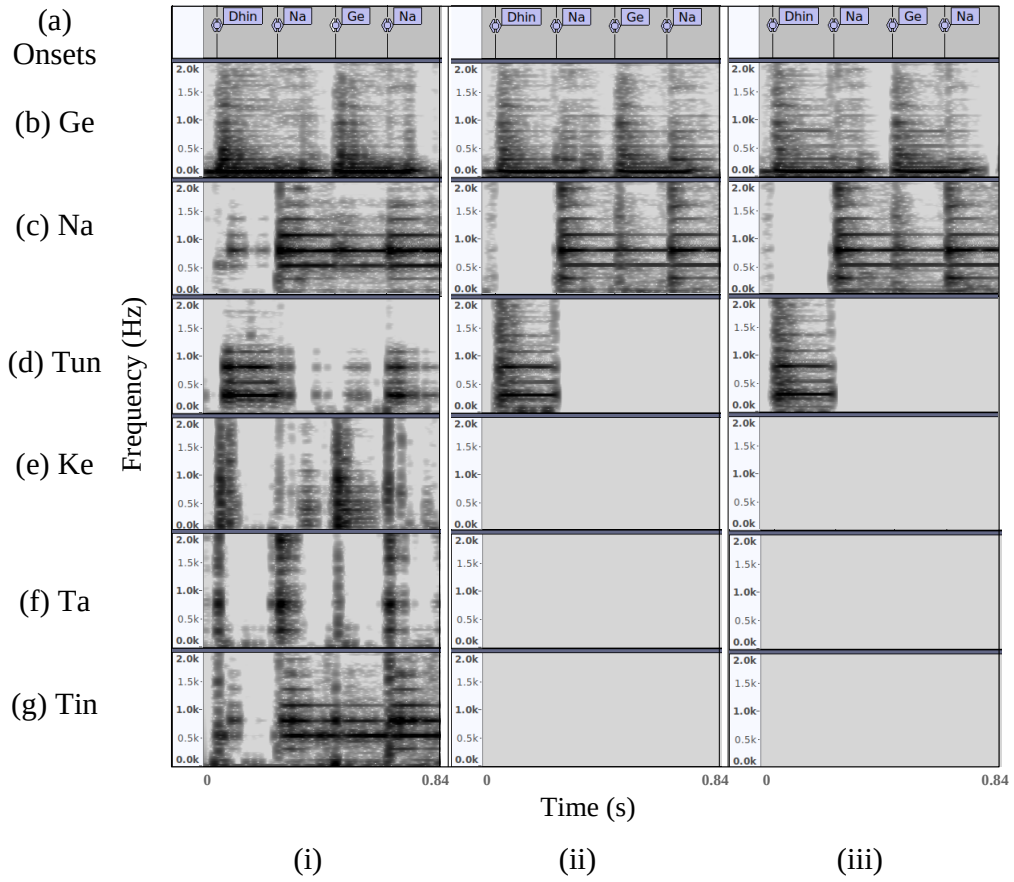


Figure 10: Magnitude spectrograms of the separated tracks of individual bols of '*Dhin Na Ge Na*'. (i) Using multiple bases without the score. (ii) Using multiple bases with additional input from the score (iii) Using single bases for onset and decay regions, with additional input from the score.

the score to further constrain the decomposition, in (ii) using the score to selectively decompose every stroke using only the required templates, and in (iii) using separate templates for onset and decay regions and also using the score like in (ii). It is evident from (i) that the onset regions of the resonant and damped strokes on both the drums are quite similar in their spectral content. This explains why the strokes *Ke* and *Ta* are also activated when *Ge* and *Na* respectively are. Also apparent is the similar harmonic structure of *Na* and *Tin*. However, as far as the interference between the and the *dayan* is concerned, fairly good separation has been achieved with the separated audio of *Na* having almost no bass frequency of the *Ge* and the *Ge* audio having almost none of the harmonic content of the *Na*. The second and third methods result in all the energy getting distributed between only the selected bols (*Na, Tun and Ge*) and hence the last three rows appear empty. But, at the cost of this zero cross-talk with the bols not present in a given frame, some interference between the and *dayan* has now crept in - the *Ge* audio has faint traces of the harmonics of *Na*, and the *Na* audio has that of the bass frequency. Between (ii) and (iii), the separation is slightly better at the onsets in (iii), as the energy in the Na audio when a *Ge* onset is present (and vice-versa) is reduced and not as pronounced as in (ii).

8 Questions for Future Work

The immediate next step is to perform intensity and pitch modulation related measurements to verify if the separation indeed helps. This is not so straightforward however, due to the lack of ground truth for the recorded data. While the same compositions have also been synthesized, additional steps to make them expressive are yet to be taken. And in order to do that, there is a need to first understand the changes that a changing intensity level causes to the spectral content of the stroke - whether a stroke being playing louder or softer are related by a mere scaling factor or if there are other changes as well.

The problem of finding note intensities to produce more complete transcriptions has been attempted on piano music with the help of parametric models to constrain the templates to have certain spectral structures, thus aiding the decomposition (Ewert and Müller, 2011). This is useful when a training step cannot be implemented due to non-availability of training data for the instruments in an unknown test audio. However, parametric models are easier to conceive for pianos due to the harmonic nature of their sounds and a lack of such a well-defined structure in broadband noise-like drum sounds has prevented parametric models from being used for western drums. In the case of the tabla however, it may be possible to constrain the templates based on such parameters, owing to the harmonic nature of its sustained sounds. This may help remove some of the harmonic interference observed even in the cases where a score was used to inform the decomposition.

References

D. S. Alves, J. Paulus, and J. Fonseca. Drum transcription from multichannel recordings with non-negative matrix factorization. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO2009)*, pages 894–898, 2009.

E. Battenberg, V. Huang, and D. Wessel. Live drum separation using probabilistic spectral clustering based on the itakura-saito divergence. In *Proceedings of the AES 45th Conference on Time-Frequency Processing in Audio*, 2012.

B. Bel and J. Kippen. Modelling music with grammars: formal language representation in the bol processor, 1992.

J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, 13(5):1035–1047, 2005.

E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F.-R. Stoter. Musical source separation: An introduction. *IEEE Signal Processing Magazine*, 36(1):31–40, 2019.

P. Chordia. Segmentation and recognition of tabla strokes. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 107–114, 2005.

P. Chordia and A. Rae. Tabla gyan: a system for realtime tabla recognition and resynthesis. In *International Computer Music Conference (ICMC)*, 2008.

D. Courtney. *Fundamentals of Tabla*. Sur Sangee Services, 2013.

D. Courtney. *Manufacture and Repair of Tabla*. Sur Sangee Services, 2016.

C. Dittmar and D. Gärtner. Real-time transcription and separation of drum recordings based on nmf decomposition. In *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, pages 187–194, 2014.

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*, 24(9):1535–1547, 2016.

J. Downing. Non-negative matrix factorization for drum source separation and transcription.

S. Ewert and M. Müller. Estimating note intensities in music recordings. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 385–388, 2011.

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

D. Fitzgerald and R. Jaiswal. On the use of masking filters in sound source separation. 2012.

O. Gillet and G. Richard. Automatic labelling of tabla signals. 2003.

S. Gupta, A. Srinivasamurthy, M. Kumar, H. A. Murthy, and X. Serra. Discovery of syllabic percussion patterns in tabla solo recordings. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 385–391, 2015.

M. Helen and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO2005)*, pages 1–4. IEEE, 2005.

A. Klapuri and M. Davy. *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems (NIPS)*, pages 556–562, 2001.

H. Lindsay-Smith, S. McDonald, and M. Sandler. Drumkit transcription via convolutive nmf. In *International Conference on Digital Audio Effects (DAFx)*, 2012.

K. Narang and P. Rao. Acoustic features for determining goodness of tabla strokes. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 257–263, 2017.

A. D. Patel and J. R. Iversen. Acoustic and perceptual comparison of speech and drum sounds in the north indian tabla tradition: An empirical study of sound symbolism. In *Proceedings of the 15th international congress of phonetic sciences (ICPhS)*, pages 925–928, 2003.

J. Paulus and T. Virtanen. Drum transcription with non-negative spectrogram factorisation. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO2005)*, pages 1–4, 2005.

A. Pradhan. *Tabla: A Performer's Perspective*. BookBaby, 2011. ISBN 9781617924248. URL `https://books.google.co.in/books?id=42laDQAAQBAJ`.

B. Ramakrishna and M. M. Sondhi. Vibrations of indian musical drums regarded as composite membranes. *The Journal of the Acoustical Society of America*, 26(4):523–529, 1954.

C. V. Raman. The indian musical drums. In *Proceedings of the Indian Academy of Sciences-Section A*, volume 1, pages 179–188, 1934.

M. A. Rohit and P. Rao. Acoustic-prosodic features of tabla bol recitation and correspondence with the tabla imitation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1229–1233, 2018.

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *International Conference on Independent Component Analysis and Signal Separation*, pages 494–499, 2004.