TRANSCRIPTION AND RECOGNITION OF SYLLABLE BASED PERCUSSION PATTERNS: THE CASE OF BEIJING OPERA

Ajay Srinivasamurthy* ajays.murthy@upf.edu
Rafael Caro Repetto* rafael.caro@upf.edu
Harshavardhan Sundar† harsha@ece.iisc.ernet.in
Xavier Serra* xavier.serra@upf.edu

*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
†Speech and Audio Group, Indian Institute of Science, Bangalore, India

ABSTRACT

In many cultures of the world, traditional percussion music uses mnemonic syllables that are representative of the timbres of instruments. These syllables are orally transmitted and often provide a language for percussion in those music cultures. Percussion patterns in these cultures thus have a well defined representation in the form of these syllables, which can be utilized in several computational percussion pattern analysis tasks. We explore a connected word speech recognition based framework that can effectively utilize the syllabic representation for automatic transcription and recognition of audio percussion patterns. In particular, we consider the case of Beijing opera and present a syllable level hidden markov model (HMM) based system for transcription and classification of percussion patterns. The encouraging classification results on a representative dataset of Beijing opera percussion patterns supports our approach and provides further insights on the utility of these syllables for computational description of percussion patterns.

1. INTRODUCTION

One common feature in traditional musics is the development of sets of predefined, identifiable melodic and rhythmic patterns. These patterns form a repository of structural elements for the composition or performance of the traditional repertoire. Certain entities of traditional music theory, like melodic modes, rhythmic cycles or musical forms, are instantiated by means of these patterns. The patterns function as key elements for the coordination of different musical elements, like instrumental and vocal sections, and the relationship with other art forms, like dance, theatrical acting, story-telling, etc. For the transmission of such patterns in these mostly oral traditions, particular systems of oral mnemonics have been developed. These systems often share common features across different cultures, so that general principles can be established. Computational analysis of these patterns is an important aspect in Music Information Research (MIR) for such music cultures. Further, their own traditional systems of transmission can offer a solid basis for their modeling.

1.1 Syllable based Percussion

Many music traditions around the world have developed particular systems of oral mnemonics for transmission of the repertoire and the technique. David Hughes [7] coined the term *acoustic-iconic mnemonic* systems for these phenomena, and described their use in different genres of traditional Japanese music. As he points out, the core aspect of these systems is that the syllables are chosen for the similarity of their phonetic features with the acoustic properties of the sounds they are representing, establishing an iconic relationship with them. Therefore, these systems are essentially different from those of solmization [6], like for instance the syllables of solfège, of the Indian svaras or the Chinese gongche notation, which are nonsensical in relation to the acoustic phenomena they represent. In this paper, we focus on the oral syllabic systems of mnemonics developed for percussion traditions.

The use of the aforementioned systems for the transmission of percussion is wide extended among traditional musics. David Hughes mentions in his paper, the *shōga* used for the set of drums of *Noh* theatre. In Korea, the young genre of *samul nori*, a percussion quartet of drums and gongs, draws on traditional syllabic mnemonics for the transmission of the repertoire. In the Indian subcontinent, both Hindustani and Carnatic music cultures have developed such oral syllabic systems of mnemonics for the percussion instruments, respectively the *bōls* in the Hindustani tradition, employed mainly by tabla players, and the *solkaṭṭu* in the Carnatic tradition, where the main percussion instrument is the mridangam. The degree of sophistication that these systems have reached in India is such that the rhythmic recitation of the syllables, which requires high skills, are commonly inserted in concerts for musical appreciation. In Carnatic music, this practice has even been consolidated into a specific music form, called *konnakōl*. Furthermore, these systems are also known to be used in Turkish traditional music and Javanese music. In this paper, we explore the use of oral syllabic system developed in the Beijing opera tradition for the computational analysis of its percussion patterns.

The benefits of using oral syllabic systems from an MIR perspective are both the cultural specificity of the approach and the accuracy of the representation of timbre, articulation and dynamics. The characterization of these percussion traditions need to consider elements that are essential to them such as the richness of their palettes of timbres, subtleties of articulation, and the different degrees and transitions of dynamics, all of which is accurately transmitted

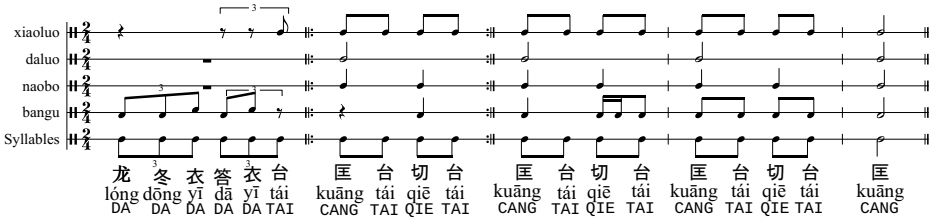

Figure 1: The percussion pattern *shanchui*. The score for each percussion instrument is shown. Shown at the bottom of the score is the syllable sequence for the score (top) and the transcription with the reduced group of syllables used in this paper (bottom). The part of the pattern enclosed between 𝄆 𝄇 can be repeated many times. The music score is only indicative, with significant variations allowed.

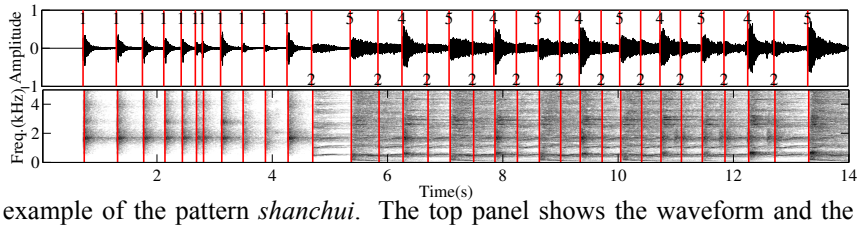

Figure 2: An audio example of the pattern *shanchui*. The top panel shows the waveform and the bottom panel is the spectrogram. The vertical lines (in red) mark the onsets of the syllables. The onsets are labeled to indicate the specific syllable group: `DA`-1, `TAI`-2, `QI`-3, `QIE`-4, and `CANG`-5 (`QI` is not present in this pattern).

by the oral syllables.

We explore the use of oral syllables as a means of representation in the MIR tasks of percussion pattern transcription and classification in syllabic percussion systems, considering Beijing opera percussion patterns as a study case. The well defined oral syllabic system and the limited set of percussion patterns make it an ideal choice for a first exploration. Since these syllables have a clear analogy to speech and language, we present a speech recognition based approach to transcribe a percussion pattern into a sequence of syllables. We then use this transcription to classify the sequence into one of the predefined set of patterns that occur in Beijing Opera. We first provide an introduction to percussion patterns in Beijing Opera.

1.2 Percussion patterns in Beijing opera

Beijing opera (Jīngjù, 京剧), also called Peking Opera, is one of the most representative genres of Chinese traditional performing arts, integrating theatrical acting with singing and instrumental accompaniment. It is an active art form and exists in the current social and cultural contexts, with a large audience and significant musicological literature. One of the main characteristics of Beijing opera aesthetics is the remarkable rhythmicity that governs the acting overall. From the stylized recitatives to the performers' movements on stage and the sequence of scenes, every element presented is integrated into an overall rhythmic flow. The main element that keeps this rhythmicity is the percussion ensemble, and the main means to fulfill this task is a set of predefined and labeled percussion patterns. Along with the main purpose of keeping the overall rhythmicity of the performance, these patterns have different functions. They signal important structural points in the play. A performance starts and ends with percussion patterns, they generally introduce and conclude arias, and mark transition points within them. They accompany the actors' movements on stage and set the mood of the play, the scene, the aria or a section of the aria. Therefore, the detection and characterization of percussion patterns is a fundamental task for the description of the music dimension in Beijing opera.

The percussion patterns in Beijing Opera music can be defined as sequences of strokes played by different combinations of the percussion instruments, and the resulting variety of timbres are transmitted using oral syllables as mnemonics. The percussion ensemble is formed mainly by five instruments played by four musicians. The *ban* clappers and the *danpigu* drum are played by one single performer, and are therefore known by a conjoint name, *bangu*. The other three instruments are the *xiaoluo* (small gong), the *daluo* (big gong) and the *naobo* (cymbals). *Bangu* has a high pitched drum-like sound while the rest of three instruments are metallophones with distinct timbres [1] . Each of the different sounds that these instruments can produce individually, either through different playing techniques or through different dynamics, as well as the sounds that are produced by a combination of different instruments have an associated syllable that represent them [9]. The syllables and their associated instrument combinations are shown in Table 1. Thus, each percussion pattern is a sequence of syllables in their pre-established order, along with their specific rhythmic structure and dynamic features. A particular feature of the oral syllabic system for Beijing opera percussion that makes it especially interesting is that the syllables that form a pattern refer to the ensemble as a whole, and not to particular instruments. Each particular pattern thus has a single unique syllabic representation shared by all the performers.

In practice, there is a library of limited set of named patterns (called luógǔ jīng, 锣鼓经) that are played in a performance, with each of these having a specific role in the arias. These named patterns in the library can be referred to as "pattern classes" for the purpose of classification, and classifying an instance of a pattern occurring in the audio recording of an aria into one of these pattern classes is thus a primary task. Although a definite agreed number for the total number of these patterns is lacking, some estimations,

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

like in [9], suggest the existence of around ninety of them.

Figure 1 shows the score for an example pattern *shanchui*. It also shows how a possible transcription in staff notation (adapted from [9]) can be simplified in a single line by the oral syllabic system. Hence, the use of these oral syllabic sequences simplify and unify the representation of these patterns played by an ensemble, making them optimal for the transcription and automatic classification of the patterns. Further, Figure 2 shows an audio example, along with time aligned markers to indicate the syllable onsets. The spectrogram shows the timbral characteristics of the percussion instruments *xiaoluo* (increasing pitch) and *daluo* (decreasing pitch). Some variation to the notated score can also be seen, such as expressive timing and additional insertion of syllables.

Though the patterns are limited in number and predefined, there are several challenges to the problem of percussion pattern transcription and classification. Being an oral tradition, the syllables used for the representation of the patterns lacks full consistency and general agreement. The result being that one particular timbre might be represented by more than one syllable. Furthermore, the syllabic representation conveys information for the conjoint timbre of the ensemble, so only the main structural sounds are represented. In an actual performance, a particular syllable might be performed by different combinations of instruments - e.g. in Figure 1, the first occurrence of the syllable tái is played just by the *xiaoluo*, but in the rest of the pattern is played by *xiaoluo* and the *bangu* together. In fact, generally speaking, the strokes of the *bangu* are seldom conveyed in the syllabic sequence (as can be seen in the third measure in Figure 1 for the second sixteenth-note of the *bangu*), except for the introductions and other structural points played by the drum alone. As indicated in Table 1, cāng is mostly a combination of all the three metallophones, but in some cases, cāng can be played with just the *daluo*, or just the *daluo*+*naobo* combination. A detailed description and scores for various patterns is available at `http://compmusic.upf.edu/bo-perc-patterns`

Since one of the main functions of the patterns is to accompany the movements of actors on stage, the overall length and the relative duration of each stroke can vary notably, which makes it difficult to set a stable pulse or a definite meter. The time signature and the measure bars used in Figure 1, as suggested in [9], are only indicative and fail to convey the rhythmic flexibility of the pattern. Furthermore, many patterns (such as *shanchui*) accompany scenic movements of undefined duration. In these cases, certain syllable sub-sequences in the pattern are repeated indefinitely, e.g. the audio example in Figure 2 has two additional repetitions of the sub-sequence cāng-tái-qiē-tái in the pattern. This causes the same pattern in different performances to have variable lengths, and these repetitions need to be explicitly handled. Finally, although the patterns are usually played in isolation, in many cases the string instruments or even the vocals can start playing before the patterns end, presenting challenges in identification and classification.

1.3 Previous work

There is significant MIR literature on percussion transcription [4]. Nakano et al. [10] explored drum pattern retrieval using vocal percussion, using an HMM based approach. They used onomatopoeia as the internal representation for drum patterns, with a focus on retrieving known fixed sequences from a library of drum patterns with snare and bass drums. Kapur et al. explored query by BeatBoxing [8], aiming to map the BeatBoxing sounds into the corresponding drum sounds. A distinction to be noted here is that in vocal percussion systems such as BeatBoxing, the vocalizations form the music itself, and not a means for transmission as in the case of oral syllables. More recently, Paulus et al. proposed the use of connected HMMs for drum transcription in polyphonic music [12]. This approach is different from what we present in the sense that it aimed to transcribe individual drums (bass, snare, hi-hat) and not overall timbres due to combinations, and no reference to syllabic percussion was made. However all these approaches have indirectly and implicitly used some form of syllabic representations for drum patterns.

Syllables	Instruments	Symbol
bā (巴, 八), běn (本), dā (答), dà (大), dōng (冬, 咚), duō (哆), lóng (龙), yī (衣)	bangu	DA
lái (来), tái (台), lìng (另)	xiaoluo	TAI
qī (七), pū (扑)	naobo	QI
qiē (切)	naobo+xiaoluo	QIE
cāng (仓), kuāng (匡), kōng (空)	daluo+<naobo> +<xiaoluo>	CANG

Table 1: Syllables used in Beijing opera percussion and their grouping used in this paper. Column 2 shows the instrument combination used to produce the syllable, instrument shown between <> is optional. Column 3 shows the symbol we use for the syllable group in this paper.

Chordia [2] explored the use of tabla *bōls* in transcription of solo tabla sequences. Recently, tabla syllables were used for a predictive model for tabla stroke sequences [3]. Anantapadmanabhan et al. [1] used the syllables of the mridangam in a stroke transcription task. Unlike these works, we address a syllabic system that conveys information for a whole ensemble instead of individual instruments.

Despite the rich musical heritage and the size of audience, little work has been done for computational analysis of Beijing opera from an MIR perspective. It has been studied as a target in a some genre classification works [17] and the acoustical properties of Beijing opera singing has been studied [14]. Apart from a recent study [15] that explored the use of Non-negative matrix factorization for onset detection and onset classification into the different percussion instrument classes, no significant work has studied Beijing opera percussion from a computational perspective.

Similar to Nakano et al. [10], we explore a speech recognition based framework in this study. This approach is different to ours in the sense that these onomatopoeic representations were created by the authors, while we are relying on already existing oral traditions. Speech recognition is a well explored research area with many state of the art algorithms and systems [5]. Hence we can apply several available tools and knowledge for computational analysis of syllabic percussion patterns. To the best of our knowl-

edge, this is the first work to explore transcription and classification of syllable based percussion patterns, as applied to Beijing opera.

2. PROBLEM FORMULATION

In Beijing opera, several syllables can be mapped to a single timbre. This many-syllable to one-timbre mapping is useful to reduce the syllable space for computational analysis of percussion patterns. We first mapped each syllable to one or several of the instrument categories considered for analysis, as explained in [15], without considering differences in playing technique or dynamics. Based on inputs from expert musicologists, we then grouped the syllables with similar timbres into five syllable groups - `DA`, `TAI`, `QI`, `QIE`, and `CANG`, as shown in Table 1. Every individual stroke of the *bangu*, both drum and clappers, have been grouped as `DA`. In the rest of the syllable groups, the *bangu* can be played simultaneously or not. The single strokes of the *xiaoluo* and the *naobo* are called `TAI` and `QI` respectively, and the combined stroke of these two instruments together is the syllable `QIE`. Finally, any stroke of the *daluo* or any combination that includes *daluo* has been notated as `CANG`. This mapping to a reduced set of syllable groups is only for the purpose of computational analysis. For the remainder of the paper, we limit ourselves to the reduced set of syllable groups and use them to represent the patterns. For convenience, when it is clear from the context, we call the syllable groups as just syllables, and denote them by the common symbol in column 3 of Table 1. Hence, in the current task, there are five syllable groups. Further, in Beijing opera, the recognition of the pattern as a whole is more important that an accurate syllabic transcription of the pattern. Due to the limited set of pattern classes and owing to all the variations possible in a pattern, we are primarily interested in classifying an audio pattern into one of the possible pattern classes. Syllabic transcription is only considered as an intermediate step towards pattern classification.

We now present a formulation for transcription and recognition of syllable based audio percussion patterns. There is a significant analogy of this task to connected word speech recognition using word models. Syllables are analogous to words and a percussion pattern to a sentence - a sequence of words. There are language rules to form a sentence using a vocabulary, just as each percussion pattern is formed with a defined sequence of syllables from a vocabulary. However unlike in the case of speech recognition where infinitely many sentences are possible, in our case we have a small number of percussion patterns to be recognized.

Consider a set of N pattern classes $\mathcal{P} = \{P_1, P_2, \cdots P_N\}$, each of which is a sequence of syllables from the set of M syllables $\mathcal{S} = \{S_1, S_2, \cdots S_M\}$. So, $P_k = [s_1, s_2, \cdots, s_{L_k}]$ where $s_i \in \mathcal{S}$ and L_k is the length of P_k. Given a test audio pattern $x[n]$, the transcription task aims to obtain a syllable sequence $P^* = [s_1, s_2, \cdots, s_{L_*}]$ and the classification task aims to assign P^* into one of the patterns in the set $\mathcal{P}$.

3. DATASET

Since there was no available dataset of Beijing opera percussion patterns, we built a representative dataset of patterns from the audio recordings of arias in the CompMusic Beijing opera research corpus [13], which is a curated collection of arias from commercially available releases spanning many different artists and recording conditions. For this study, we chose only the patterns at the beginning of the aria, which are characteristic and important. From all the pattern classes existing in the corpus, we chose five ($N = 5$) most frequently used ones. These five patterns are also the most widely used and hence hold a high degree of representativeness. The patterns were extracted from the audio recording and assigned to a pattern class by a musicologist. The dataset is described in Table 2 and comprises about 22 minutes of audio with over 2200 syllables in total. The audio samples are stereo recordings sampled at 44.1kHz. The syllabic transcription of each audio pattern is obtained directly from the score of the pattern class it belonged to. Hence the ground truth transcriptions available in the dataset are not time aligned. Since it is a significant effort to obtain time aligned transcriptions, we aim to develop algorithms which do not require the use of time-aligned transcriptions for training. This also ensures that the approaches scale when we add more pattern classes to the dataset. In case of patterns where a sub-sequence of the pattern can be repeated (e.g. *man changchui* and *shanchui*), the additional syllables that occur due to repetitions were manually added by listening to the pattern. Though most of the dataset consists of isolated percussion patterns, there are many audio examples that contain a melodic background apart from the percussion pattern. The dataset is available for research purposes through a central online repository [2] .

Pattern Class	ID	Instances	$\overline{LEN}$ (σ)
daobantou 【导板头】	1	66	8.70 (1.73)
man changchui 【慢长锤】	2	33	13.99 (4.47)
duotou 【夺头】	3	19	7.18 (1.49)
xiaoluo duotou【小锣夺头】	4	11	8.16 (2.15)
shanchui【闪锤】	5	8	10.31 (3.26)
Total		**133**	**9.85 (3.69)**

Table 2: The Beijing Opera Percussion Pattern (BOPP) dataset. The last column is the mean pattern length and standard deviation in seconds.

4. THE APPROACH

The syllables are non-stationary signals and to model their timbral dynamics, we build an HMM for each syllable (analogous to a word-HMM). Using these syllable HMMs and a language model, an input audio pattern is transcribed into a sequence of syllables using Viterbi decoding, and then classified to a pattern class in the library using a measure of distance.

A block diagram of the approach is shown in Figure 3. We first build syllable level HMMs $\{\lambda_m\}, 1 \leq m \leq M(= 5)$, for each syllable S_m using features extracted from the training audio patterns. We use the MFCC features to model the timbre of the syllables. To capture the temporal dynamics of syllables, we add the velocity and the acceleration coefficients of the MFCC. The stereo audio is converted to mono, since there is no additional information in stereo channels. The 13 dimensional (including the 0^{th}

[2] More details at `http://compmusic.upf.edu/bopp-dataset`

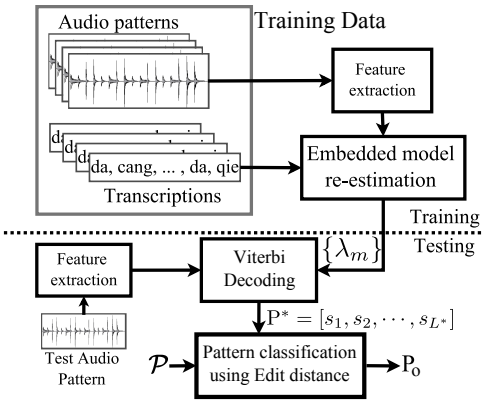

Figure 3: The block diagram of the approach

coefficient) MFCC features are computed from audio patterns with a frame size of 23.2 ms and a shift of 5.8 ms. We also explore the use of energy (as measured by the 0th MFCC coefficient) in classification performance. Hence we have two sets of features, MFCC_0_D_A, the 39 dimensional feature including the 0th, delta and double-delta coefficients, and MFCC_D_A, the 36 dimensional vector without the 0th coefficient.

We model each syllable using a 5-state left-to-right HMM including an entry and an exit non-emitting states. The emission densities for each state is modeled with a four component Gaussian Mixture Model (GMM) to capture the timbral variability in syllables. We experimented with eight and sixteen component GMMs, but with little performance improvement. Since we do not have time aligned transcriptions, an isolated HMM training for each syllable is not possible. Hence we use an embedded model Baum-Welch re-estimation to train the HMMs using just the syllable sequence corresponding to each feature sequence. The HMMs are initialized with a flat start using all of the training data. All the experiments were done using the HMM Toolkit (HTK) [16].

For testing, since we only need a rough syllabic transcription independent of the pattern class, we treat the test pattern as a first order time-homogenous discrete Markov chain, which can consist of any finite length sequence of syllables, with uniform unigram and bi-gram (transition) probabilities, i.e. $p(s_1 = S_i) = 1/M$ and $p(s_{k+1} = S_j/s_k = S_i) = 1/M, 1 \leq i, j \leq M$ and with k being the sequence index. This also forms the language model for forming the percussion patterns using syllables. Given the feature sequence extracted from test audio pattern, we use the HMMs $\{\lambda_m\}$ to do a Viterbi (forced) alignment, which aims to provide the best sequence of syllables P^*, given a syllable network constructed from the language model.

Given the decoded syllable sequence P^*, we compute the string edit distance [11] between P^* and elements in the set $\mathcal{P}$. The use of edit distance is motivated by two factors. First, due to errors in Viterbi alignment, P^* can have insertions (I), deletions (D), substitutions (B), and transposition (T) of syllables compared to the ground truth. Secondly, to handle the allowed variations in patterns, an edit distance is preferred over an exact match to the sequences in $\mathcal{P}$. We explore the use of two different string edit distance measures, Levenshtein distance (d_1) that considers I, D, B errors and the Damerau–Levenshtein distance (d_2) that considers I, D, B, T errors.

Feature	Syllable		Pattern	
	C	A	d_1	d_2
MFCC_D_A	78.14	26.32	93.23	89.47
MFCC_0_D_A	84.98	39.63	91.73	89.47

Table 3: Syllable transcription and Pattern classification performance, with Correctness (C) and Accuracy (A) measures for syllable transcription. Pattern classification results are shown for both distance measures d_1 and d_2. All values are in percentage.

As discussed earlier, there can be repetitions of a subsequence in some patterns. Though the number of repetitions is indefinite, we observed in the dataset that there are at most two repetitions in a majority of pattern instances. Hence for the pattern classes that allow repetition of a subsequence, we compute the edit distance for the cases of zero, one and two repetitions and then take the minimum distance obtained among the three cases. This way, we can handle repeated parts in a pattern. Finally, the P^* is assigned to the pattern class $P_o \in \mathcal{P}$ for which the edit distance d (either d_1 or d_2) is minimum, as in Eqn 1.

$$P_o = \underset{1 \leq k \leq N}{\operatorname{argmin}}\, d(P^*, P_k) \tag{1}$$

5. RESULTS AND DISCUSSION

We present the syllable transcription and pattern classification results on the dataset described in Section 3. The results shown in Table 3 are the mean values in a leave-one-out cross validation. We report the syllable transcription performance using the measures of Correctness (C) and Accuracy (A). If L is the length of the ground truth sequence, C=$(L-D-B)/L$ and A=$(L-D-B-I)/L$. The Correctness measure penalizes deletions and substitutions, while Accuracy measure additionally penalizes insertions too. The pattern classification performance is shown for both edit distance measures d_1 and d_2 in Table 3. All the results are reported for both the features, MFCC_0_D_A and MFCC_D_A. The difference in performance between the two features was found to be statistically significant for both Correctness and Accuracy measures in a Mann-Whitney U test at $p = 0.05$, assuming an asymptotic normal distribution.

In general, we see a good pattern classification performance while syllable transcription accuracy is poor. We see that MFCC_0_D_A has a better performance with syllable transcription, while both kinds of features provide a comparable performance for pattern classification. Though syllable transcription is not the primary task we focus on, an analysis of its performance provides several insights. The set of percussion instruments in Beijing Opera is fixed, but there can be slight variations across different instruments of the same kind. The training examples are varied and representative, and models built can be presumed to be source independent. Nevertheless, there can be unrepresented syllable timbres in test data leading to a poorer transcription performance. A bigger training dataset can improve the performance in such a case. The energy co-efficient provides significant information about the kind of syllables

ID	1	2	3	4	5	Total
1	100					62
2		93.9			6.1	33
3	10.5		68.4		21.1	19
4			18.2	81.8		11
5		12.5			87.5	8

Table 4: The confusion matrix for pattern classification, using the feature MFCC_0_D_A with d_1 distance measure. The rows and column headers represent the True Class and Assigned Class, respectively. Class labels correspond to the ID in Table 2. The last column shows the total examples in each class. All other values are in percentage and the empty blocks are zeros (omitted for clarity).

and hence gives a better syllable transcription performance.

We see that the Correctness is higher than Accuracy showing that the exact sequence of syllables, as indicated in the score was never achieved in a majority of the cases, with several insertion errors. This is due to the combined effect of errors in decoding and allowed variations in patterns. An edit distance based distance measure for classification is quite robust in the present five class problem and provides a good classification performance, despite the low transcription accuracy. Both distance measures provide comparable performance, indicating that the number of transposition errors are low. To see if there are any systematic classification errors, we build a confusion matrix (Table 4) with one of the well performing configurations: MFCC_0_D_A with d_1 distance. We see that *duotou* has a low recall, and gets confused with *shanchui* (ID=5) often. A close examination of the scores showed that a part of the pattern *duotou* is contained within *shanchui*, which explains source of confusion. Such confusions can be handled with better language models, which need further exploration.

6. CONCLUSIONS AND SUMMARY

We presented a formulation based on connected-word speech recognition for transcription and classification of syllabic percussion patterns. On a representative collection of Beijing opera percussion patterns, the presented approach provides a good classification performance, despite a simplistic language model and inadequate syllabic transcription accuracy. Though the approach is promising, the evaluation using a small dataset necessitates a further assessment of the generalization capabilities of the proposed approach. We intend to explore better language models that use sequence and rhythmic information more effectively, and extend the task to a much larger dataset spanning more pattern classes. We used isolated patterns in this study, but an automatic segmentation of patterns from audio is a good direction for future work. We also plan to extend this formulation for computational description of percussion patterns in other music cultures such as Hindustani and Carnatic music, which have more complex syllabic percussion systems.

Acknowledgments

This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as a part of the CompMusic project (ERC grant agreement 267583)

7. REFERENCES

[1] A. Anantapadmanabhan, A. Bellur, and H. A. Murthy. Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorization. In *Proc. 38th IEEE Proc. Int'l Conf. on Acoust., Speech, and Signal Processing*, pages 181–185, Vancouver, Canada, May 2013.

[2] P. Chordia. *Automatic Transcription of Solo Tabla Music*. PhD thesis, Stanford University, 2005.

[3] P. Chordia, A. Sastry, and S. Senturk. Predictive Tabla Modelling Using Variable length Markov and Hidden Markov Models. *Journal of New Music Research*, 40(2):105–118, 2011.

[4] D. FitzGerald and J. Paulus. Unpitched Percussion Transcription. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music Transcription*, pages 131–162. Springer US, 2006.

[5] X. Huang and L. Deng. An overview of modern speech recognition. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing*, Chapman & Hall/CRC Machine Learning & Pattern Recognition, pages 339–366. Chapman and Hall/CRC, 2nd edition, February 2010.

[6] A. Hughes and E. Gerson-Kiwi. Solmization. In *Grove music online. Oxford Music Online*. Oxford University Press, accessed July 18, 2014.

[7] D. Hughes. No nonsense: the logic and power of acoustic-iconic mnemonic systems. *British Journal of Ethnomusicology*, 9(2):93–120, 2000.

[8] A. Kapur, M. Benning, and G. Tzanetakis. Query by beatboxing: Music information retrieval for the dj. In *Proc. of the 5th Int'l Conf. on Music Information Retrieval (ISMIR)*, Barcelona, Spain, October 2004.

[9] W. Mu(穆文义). *Jingju dajiyue jiqiao yu lianxi: yanzou jiaocheng* 京剧打击乐技巧与练习：演奏教程 *(Technique and practice of Beijing opera percussion music: a performance course)*. Renmin yinyue chubanshe, Beijing, 2007.

[10] T. Nakano, J. Ogata, M. Goto, and Y. Hiraga. A drum pattern retrieval method by voice percussion. In *Proc. of the 5th Int'l Conf. on Music Information Retrieval (ISMIR)*, pages 550–553, October 2004.

[11] G. Navarro. A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, 33(1):31–88, March 2001.

[12] J. Paulus and A. Klapuri. Drum Sound Detection in Polyphonic Music with Hidden Markov Models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009(497292):1–9, 2009.

[13] X. Serra. Creating research corpora for the computational study of music: the case of the compmusic project. In *Proc. of the 53rd AES Int'l Conf. on Semantic Audio*, London, January 2014.

[14] J. Sundberg, L. Gu, Q. Huang, and P. Huang. Acoustical study of classical Peking Opera singing. *Journal of Voice*, 26(2):137–143, March 2012.

[15] M. Tian, A. Srinivasamurthy, M. Sandler, and X. Serra. A study of instrument-wise onset detection in beijing opera percussion ensembles. In *Proc. 39th IEEE Proc. Int'l Conf. on Acoust., Speech, and Signal Processing*, pages 2174–2178, Florence, Italy, May 2014.

[16] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.

[17] Y. Zhang and J. Zhou. A study on content-based music classification. In *Proc. of the Seventh Int'l Symp. on Signal Processing and its Applications*, volume 2, pages 113–116, Paris, France, July 2003.