Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004),
pp.550-553, October 2004.

A DRUM PATTERN RETRIEVAL METHOD BY VOICE PERCUSSION

Tomoyasu Nakano[†] *Jun Ogata*[‡] *Masataka Goto*[‡] *Yuzuru Hiraga*[†]
[†]Graduate School of Library, Information and Media Studies
University of Tsukuba, Japan
[‡]National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper presents a method for voice percussion recognition and its application to drum pattern retrieval. Recognition of voice percussion (verbalized expression of drum sound by voice) requires an approach that is different from existing methods. Individual differences in both vocal characteristics and the kinds of verbal expressions used add further complication to the task. The approach taken in this study uses onomatopoeia as internal representation of drum sounds, and combines the recognition of voice percussion with the retrieval of intended drum patterns. This scheme is intended to deal with the two types of individual differences mentioned above. In a recognition experiment with 200 utterances of voice percussion, our method achieved a recognition rate of 91.0% for the highest-tuned setting.

keywords: voice percussion recognition, drum pattern retrieval, onomatopoeia representation

1. INTRODUCTION

This paper presents a method for voice percussion recognition and its application to drum pattern retrieval. *Voice percussion* in our context is the mimicking of drum sounds by voice, expressed in verbal form that can be transcribed into phonemic representation, or *onomatopoeia* (*e.g. don-don, ta-ta*). In this sense, voice percussion is not a direct, faithful reproduction of the acoustic properties of actual drum sounds.

The analysis and recognition of voice percussion requires an approach that is different from those in existing work, such as drum sound recognition [1, 4] or query-by-humming [6]. Drum sound recognition looks for acoustic properties that are characteristic of the instrument, but in our case, mapping between the input voice and target instrument sound is only indirect and metaphoric. Query-by-humming focuses on pitch detection and melodic feature extraction, but these have less relevance in voice percussion recognition, which is primarily concerned with classification of timbre and identification of articulation methods. Differences among individuals in both vocal characteristics and the kinds of verbal expressions used add further complication to the task.

The following sections describe our approach and the experimental results of its evaluation. Section 2 presents an overview of the proposed method. Section 3 describes the experiments for deriving the model and its evaluation. Section 4 concludes the paper, with discussion on provisions for future work.

We view voice percussion recognition to have promising applications in widening the scope of music information retrieval methods, and introducing new possibilities for music transcription, composition and arrangement.

2. PROPOSED METHOD

The proposed method combines voice percussion recognition with drum pattern retrieval. A *drum pattern* is a sequence of percussion beats typically used in popular music (see Figure 2 for example drum patterns). In the current work, drum patterns consist of two percussion instruments — bass drum (BD) and snare drum (SD). A voice percussion input is regarded as expressing a particular drum pattern. So instead of identifying individual drum sounds in isolation, the entire sequence is matched with the entries of a drum pattern database. The recognition is considered to be successful if the intended pattern is correctly retrieved. An overview of the method is shown in Figure 1.

Reflecting the verbal nature of the input, onomatopoeia is used as an intermediate level representation of voice percussion (Fig.1, B2, B3). This plays the pivotal role of connecting the recognition and retrieval phases.

Typical onomatopoeic expressions of single drum sounds are stored in a pronunciation dictionary (B2), where the entries are derived from the results of existing work on onomatopoeia [5], and also from the extraction experiment described in 3.1. Using this dictionary, the entries of the drum pattern database (D) are expanded into possible onomatopoeic expressions (B3).

The elements of these expressions are mapped into sequences of acoustic features by an *acoustic model* (B4), which is a stochastic transitional network using the Hidden Markov Model (HMM). The base acoustic model is a phoneme-level HMM (monophone) provided by CSRC [7], trained by a newspaper dictation corpus (40,000 utterances by 270 Japanese speakers).

For a given voice percussion utterance (A1), our method

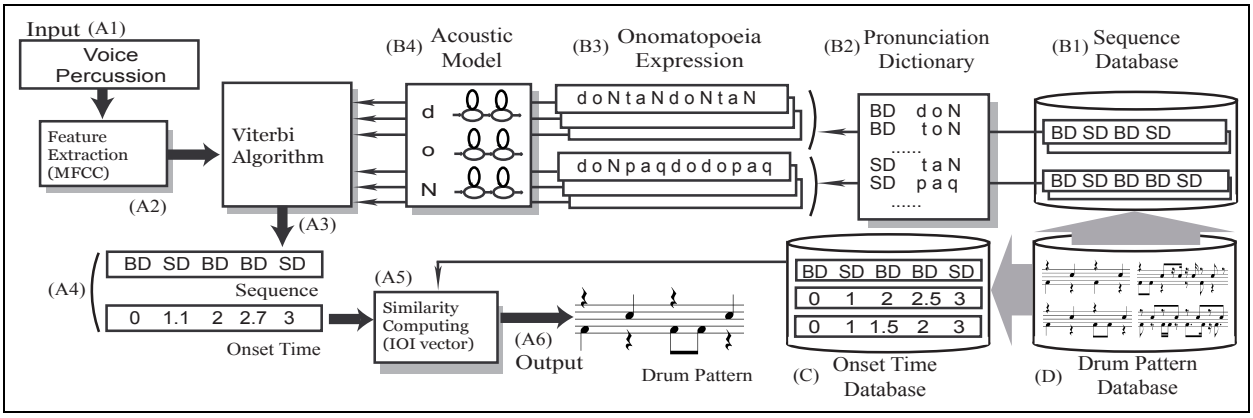

Figure 1. Overview of the proposed method.

first extracts its acoustic features in the form of MFCCs (Mel Frequency Cepstral Coefficients: A2). This output is matched with the acoustic model using the Viterbi algorithm (A3). The *instrument name sequence* with the highest likelihood score is selected as the retrieved result (A4). This stage of retrieval, where only the instrument name sequence is retrieved, is called *sequence matching*.

A single instrument name sequence may correspond to several drum patterns with different rhythm. *Drum pattern matching* further matches the *IOI (Inter-Onset Interval) vector* with entries in the onset time database (C). Here, the IOI vector is defined as the sequence of temporal differences between adjacent onset times. The similarity of the IOI vectors is calculated by the *cosine measure*, and the drum pattern with the highest score is selected as the retrieved result (A6).

The acoustic model and the pronunciation dictionary can be switched from their basic versions to more tuned versions (*e.g.* reflecting individual difference), providing various settings for the recognition experiments.

3. EXPERIMENTS AND RESULTS

The proposed method is implemented and tested in two kinds of experiments. Experiment 1 is a preliminary psychological experiment of *expression extraction*, which gathers data on how subjects express drum sounds and patterns. Experiment 2 is a series of *recognition* experiments, which evaluates the performance of the proposed method under various settings.

The target drum patterns have the length of 4 beats (one measure in 4/4 time), and consist of Bass Drum (BD) and Snare Drum (SD) sounds with no simultaneous beats.

3.1. Experiment 1: Expression Extraction

The purpose of this experiment is to identify and extract typical onomatopoeic expressions of drum patterns by human subjects. The data obtained in the experiment are used for the construction and tuning of the pronunciation dictionary and the acoustic model. The data are also used as test data for the recognition experiments described in 3.2.

3.1.1. Experiment Setting and Task

The subjects were instructed to listen to ten kinds of drum patterns (each in 80 and 120 M.M. tempo), and to freely sing out each pattern according to their verbal image. Figure 2 shows the drum patterns used in the experiment. The lower notes correspond to the BD sounds, and the higher notes, the SD sounds. The sound stimuli were generated by using the BD and SD sounds in the RWC Music Database (Musical Instrument Sound) [3] [1] .

There were 17 subjects (all Japanese speakers) with ages from 19 to 31. Two had previous experience with percussion instruments. The 20 patterns were presented to each subject in random order.

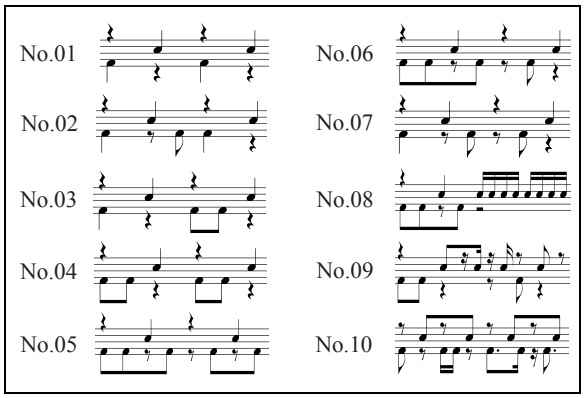

Figure 2. Drum patterns used in the extraction experiment.

3.1.2. Results and Discussion

The onomatopoeic expressions for individual drum sounds could be classified into six types — namely "CV", "CVQ", "CVN", "CVNQ", "CVRN" and "CVRNQ". The symbols used above stand for:

C: consonant (t, d, p, ...)
V: vowel (a, o, e, ...)
Q: choked sound, written as *q*
N: syllabic nasal (*n* sound)
R: long vowel

[1] 421BD1N3.WAV, 421SD3N3.WAV (RWC-MDB-I-2001 No.42)

For example, with “t” for C and “a” for V, CV, CVQ and CVN are “ta”, “taq”, and “tan”, respectively.

The leading CV pair was generally significant for instrument identification (between BD and SD). Typical CV pairs used for BD and SD are shown in Table 1, together with the number of subjects that used them (most subjects used several CV pairs for the same instrument). The trailing parts (of Q, R, N) had correspondence with the length and rhythmic patterns of the beats.

	BD						
type	/d o/	/d u/	/t o/	/t u/	/z u/	/r e/	
subjects	9	8	5	4	3	1	
	SD						
type	/t a/	/d a/	/p a/	/k a/	/ch a/	/t e/	/r a/
subjects	13	3	3	2	1	1	1

Table 1. Examples of CV pairs used for drum sounds.

Although most subjects distinguished timbre by onomatopoeic expression, two subjects made the BD/SD distinction mostly by pitch. These two cases were excluded from the recognition experiment described below. Another two subjects (both with percussion instrument experience) verbally expressed rest notes, which may reflect their ordinary practice.

But in general, the obtained data seemed to ensure the validity of the proposed method, which is tested in experiment 2.

3.2. Experiment 2: Recognition Experiments

The proposed method was tested in a series of recognition experiments. The drum pattern database used in the experiments had 1169 drum patterns (538 instances of instrument sequence patterns), including all ten patterns used in the extraction experiment (Fig.2). 1167 of the drum patterns were extracted from the SMF (Standard MIDI File) data of the RWC Music Database (Popular Music) [2], and the remaining two are patterns from the extraction experiment not included in the SMF data (patterns 8 & 10).

The recognition experiments used the utterance data of 15 subjects in the extraction experiment, which are divided into two groups — test set (10 subjects, 200 utterances) and training set (5 subjects, 100 utterances).

3.2.1. Experimental Setting

Four experiments (A)–(D) were performed, over a combination of three acoustic models and two pronunciation dictionaries. The four conditions are as follows.

(A) Base acoustic model of Japanese phonemes (model 1), basic pronunciation dictionary of Japanese phonemes.
(B) Speaker-independent acoustic model (model 2), basic dictionary.
(C) Speaker-specific acoustic model (model 3), basic dictionary.
(D) Acoustic model 3, pronunciation dictionary with entries restricted to those used in voice percussion utterances (individual dictionary).

Models 2 & 3 were tuned by using MLLR-MAP [8] as the adaptation method, which is a combination of MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum *A Posteriori* Probability). Model 2 is tuned with the training set, corresponding to a general acoustic model of voice percussion. Model 3 is tuned with the utterance data of individual subjects in the test set.

3.2.2. Evaluation Criteria

The results were evaluated by three criteria, namely, sequence evaluation, onset evaluation, and drum pattern evaluation.

Sequence evaluation is the recognition rate (percentage of test data correctly recognized) of instrument name sequences; where the recognized sequence is the one with the highest likelihood score obtained by the Viterbi algorithm. *Onset evaluation* checks the deviation of extracted onsets from the correct onsets (hand-labeled). *Drum pattern evaluation* is the recognition rate of drum patterns, *i.e.* combination of instrument name and onset sequences. The onset sequence (*IOI vector*) having the highest similarity (*cosine measure*) with the input is selected as the recognized result. Since different drum patterns may have the same instrument name sequence, this is a more strict criteria than sequence evaluation.

The recognition rates for acoustic model 3 are evaluated by using a cross validation method. In each trial, the model is tuned by 18 of the 20 utterances by a single subject, and is tested by using the remaining two as test data. The overall recognition rate is the result of 10 trials per subject for the 10 subjects in the test set.

3.2.3. Results and Discussion

Table 2 shows the recognition rates for sequence evaluation and drum pattern evaluation. The results show an increasing trend with a large improvement (of over 15%) between experiments (B) and (C) — indicating that incorporating task-specific and individually tuned data is effective for improving system performance.

Figures 3 and 4 show the sequence evaluation results by subject and by drum pattern. The increasing trend is especially notable for subjects I and IV, where the recognition rates improved from 0% in experiment (A) to 70–95%

Condition	Acoustic Model	Pronunciation Dictionary	Recognition Rate	
			Sequence	Drum Pattern
(A)	model **1**	basic	65.0%	62.5%
(B)	model **2**	basic	70.0%	68.5%
(C)	model **3**	basic	89.0%	86.5%
(D)	model **3**	individual	93.0%	91.0%

Table 2. Experimental conditions and results.

in experiment (D). This is the main source of the overall improvement between (B) and (C).

Figure 5 shows the onset evaluation results for experiment (D). The overall difference is small (mean=-0.0220 sec, standard deviation=0.0231 sec), indicating that onset extraction is precise enough in general. Larger deviations occurred in cases such as when a silent pause between adjacent sounds was not properly detected, two sounds could not be separated, or noise such as breath sound was falsely detected.

The precision of onset detection is also reflected in the relatively small difference between sequence evaluation and drum pattern evaluation in Table 2. This suggests that the correct recognition of instrument name sequences comes hand-in-hand with the correct recognition of the onset time.

Although recognition failure in experiments (A) and (B) are mainly due to lack of task-specific data and insufficient tuning, there are still failures in the more precise settings of experiments (C) and (D). These failures are due to cases such as when rests were spoken out to maintain rhythm, or when instrument difference was expressed by pitch. Dealing with these and other problems are issues of future work.

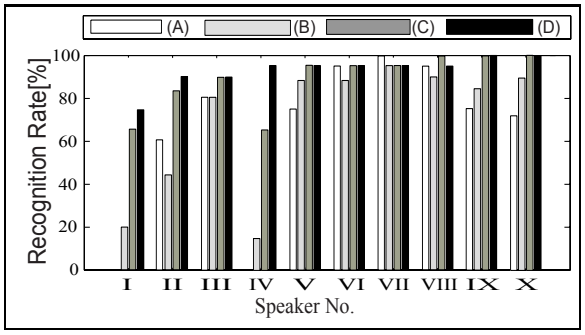

Figure 3. Result of sequence evaluation (by subject).

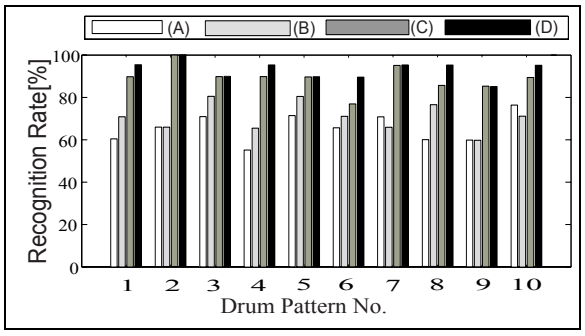

Figure 4. Result of sequence evaluation (by drum pattern).

4. GENERAL DISCUSSION AND CONCLUSION

This paper presented a method for voice percussion recognition and drum pattern retrieval. The results indicate the general effectiveness of our method, obtaining a recognition rate of about 70% as the bottom line. The increase of the recognition rate for tuned settings (experiments (C) and (D)) suggests that individual difference is wide ranged, and from a practical point of view, system performance can be improved dramatically by adjusting to individual users.

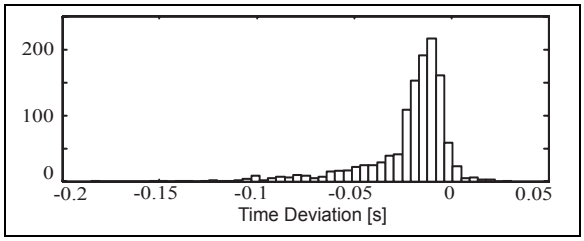

Figure 5. Result of onset evaluation.

Still, the current framework is restricted in important aspects. First, the data used is obtained from Japanese speakers, so the system should be seen to be adjusted to Japanese users. Whether the general framework of our method can be applicable to users of different languages and background culture needs investigation.

Second, the current model handles only two instruments, namely SD and BD. Extending the scope of instruments (*e.g.* including tams, cymbals, etc.), and the complexity of drum patterns (*e.g.* allowing simultaneous beats) is an immediate issue for future work.

There are many prospective applications conceivable for our method. The most direct application is in music information retrieval, where the scope of retrieval will be greatly enhanced by dealing with percussion sounds and rhythmic patterns. Other possible application can be found in composition and performance supporting systems. The current work is intended to be the first step toward opening such new possibilities.

5. REFERENCES

[1] Goto, M. et al. "A Sound Source Separation System for Percussion Instruments", *IEICE Transactions*, J77-D-II, pp.901-911, 1994 (*in Japanese*).

[2] Goto, M. et al. "RWC Music Database: Popular, Classical, and Jazz Music Databases", *Proc. of ISMIR*, pp.287-288, 2002.

[3] Goto, M. et al. "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", *Proc. of ISMIR*, pp.229-230, 2003.

[4] Herrera, P. et al. "Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques", *Proc. of ICMAI*, LNAI2445, pp.69-80, 2002.

[5] Hiyane, K. et al. "Study of Spectrum Structure of Short-time Sounds and its Onomatopoeia Expression", *Technical Report of IEICE. SP97-125*, pp.65-72, 1998 (*in Japanese*).

[6] Kageyama, T. et al. "Melody Retrieval with Humming", *Proc. of ICMC*, pp.349-351, 1993.

[7] Lee, A. et al. "Continuous Speech Recognition Consortium — an open repository for CSR tools and models —", *In Proc. IEEE Int'l Conf. on Language Resources and Evaluation (LREC2002)*, pp.1438-1441, 2002.

[8] Thelen, E. et al. "Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition", *Proc. of ICASSP*, pp.1035-1038, 1997.