Beatbox sounds recognition using a speech-dedicated HMM-GMM based system

Solène Evain, Adrien Contesse, Antoine Pinchaud, Didier Schwab, Benjamin Lecouteux, Nathalie Henrich Bernardoni

BEATBOX SOUNDS RECOGNITION USING A SPEECH-DEDICATED HMM-GMM BASED SYSTEM

Solène Evain[1], Adrien Contesse[2*], Antoine Pinchaud[*], Didier Schwab[1], Benjamin Lecouteux[1], and Nathalie Henrich Bernardoni[3]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[*]http://www.vocalgrammatics.fr/
[2]ÉSAD Amiens, De-sign-e Lab, 80080 Amiens, France
[3]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

solene.evain@univ-grenoble-alpes.fr, AdrienContesse@gmail.com, APinchaud@gmail.com, Didier.Schwab@imag.fr, Benjamin.Lecouteux@imag.fr, nathalie.henrich@gipsa-lab.fr

Abstract: **Human beatboxing is a vocal art making use of speech organs to produce percussive sounds and imitate musical instruments. Beatbox sounds classification is a current challenge. We propose a beatbox sounds recognition system with an adaptation of the Kaldi toolbox, widely used for automatic speech recognition (ASR). Our corpus is composed of isolated sounds produced by two beatboxers and is composed of 80 different sounds. We focused on decoding with monophones acoustic models, trained with a HMM-GMM model. One type of transcription was used: a beatbox specific writing system named Vocal Grammatics (VG) which uses concepts of articulatory phonetics.**

Keywords: Human beatbox, automatic speech recognition, Kaldi, isolated sounds recognition

I - INTRODUCTION

Human beatboxing emerged in the '80s in the Bronx, a borough of New York City, and is associated with hip-hop culture. It consists in producing vocal percussions along with musical instruments imitations, such as trumpet or guitar. Beatbox sounds classification can be used for Music Information Retrieval as a request for searching different types of music [4] or for voice-controlled applications with a user-defined number of classes [3]. Good classification rates were obtained with an ACE-based system[1] on a limited range of classes, i.e. five main beatbox sounds *bass drum*, *open hi-hat*, *closed hi-hat*, *k-snare* and *p-snare* drums [7]. To the best of our knowledge, automatic recognition of beatbox sounds using a speech recognition system has only been explored by [5]. Their training database consists of isolated drum beatbox sounds (five classes *cymbal*, *hi-hat*, *kick*, *rimshot* and *snare*) and instrument imitations (8 classes). Performance was poor for instrument imitations (best recognition error rate of 41%), yet good performance was demonstrated for limited beatbox sounds classes (best recognition error rate of 9%).

In continuing this effort towards the development of an efficient and reliable automatic beatbox sounds recognition system, we aim to extend the number of sound classes and enable the recognition of subtle variants in beatbox sounds production. We consider human beatbox as a musical language composed of sound units that we shall call *boxemes* with reference to speech phonemes. This work was made with a view of creating an interactive artistic setup that would provide visual feedbacks during beatbox sounds production.

The paper is structured as follows. Section II presents the training database. The recognition system is presented in Section III. Different experiments are described in Section IV and their results given in Section V. Sections VI and VII provide a discussion and conclusion of the paper, along with future works.

II - MATERIALS

Our beatbox sounds corpus was recorded by two male beatboxers: a professional beatboxer (third author, stage name *Andro*) and an amateur one (second author). It is composed of 80 boxemes and could be considered as a large vocabulary corpus compared to previous corpora used in papers for classification. Isolated sounds only are considered here, rhythmic sequences being discarded in

[1]Autonomous Classification Engine or ACE, developed for optimising music classification

first approach.

A articulatory-based pictographic writing system developed by second author and called *Vocal Grammatics* [1] was used for annotation. In this latter, the glyphs are composed of two pieces of information: one about the speech organs that are used, and one about the manner the sounds are produced (plosive, fricative...). Fig. 1 illustrates this writing system in the case of a bilabial plosive sound with a morphological glyph representing two lips and a symbolic cross-shaped glyph representing plosion.

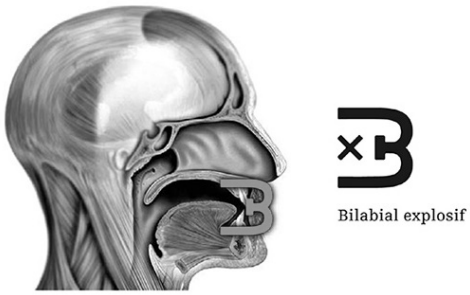


Figure 1: Representation of a bilabial plosive sound with *Vocal Grammatics* pictographic writing system

Our Large Vocabulary Corpus was recorded with six microphones. Five of them were recording simultaneously and one was encapsulated (one or two hands cover the capsule of the microphone). The microphones differed in terms of specificities (e.g. condenser vs dynamic) and placement. Table 1 give the details of the microphones when table 2 is a recap chart of the composition of the corpus.

Microphone	**Distance from the mouth**	**Specificity**
Brauner VM1 (braun)	10 cm	condenser + pop filter
DPA 4006 (ambia)	50 cm	condenser ambient mic
DPA 4060 (tie)	10 cm	condenser
Shure SM58 (sm58p)	10 cm	dynamic
Shure SM58 (sm58l)	15 cm	dynamic
Shure beta 58 (beta)	1 cm	dynamic + encapsulated

Table 1: Recap chart of the different microphones

	Large Voc. Corp.
Beatboxers	Adrien (amateur), Andro (professional)
Date	2019
Num. of sounds	80
Num. of sounds per beatboxer	Adrien: 56/80 Andro: 80/80
Transcription	Vocal Grammatics
Microphone	5 simultaneous + 1 encapsulated
Recording parameters	44100 Hz, 16 bits, mono, wav
Train	
Recording time	∼92mn
Repetitions	6 or 2
Test	
Recording time	∼114mn
Repetitions	7 (on average)

Table 2: Recap chart of the corpus

The training and testing of acoustic models was made with the Kaldi toolbox [6].

III - Speech recognition

Our approach is based on the assumption that human beatbox is structured like a musical language, using the speech organs to produce sound units that can be distinguished from each other and that each have a specific musical meaning for the beatboxer. In this context, a speech-dedicated recognition system could make it possible to automatically recognise beatbox productions. An ASR system can either be word-based or sub-word based [2]. Word-based systems require a model for each word in the vocabulary, trained with many repetitions of each word which are supposed to give the system a representation of the variability in speech production. Instead of cutting a sound into sub-words units and have a model per phoneme, a word-based system focuses on the sound as a whole and recognises the words it has been trained on only. This preliminary work focuses on isolated words recognition. Co-articulation or word boundary were discarded, yet keeping constraints on noise treatment, intra- and inter-speaker variability.

Mel Frequency Cepstral Coefficient (MFCC) acoustic features were extracted. They are based on human peripheral auditory system ([8]) and are widely used in ASR. Each beatbox sound was associated with a Hidden Markov Model (HMM).

IV - METHODS

Several systems were trained for the purpose of testing different recognition parameters. The influence of using different microphones with different placements and sensitivity was studied in order to know whether all the recordings could be used together for training a robust system.

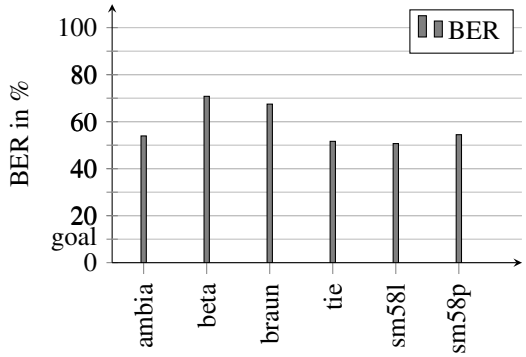
Finally, different parameters were tested: the increase of HMM states, the increase of MFCC, the addition of a pause phoneme in the lexicon and the increase of the silence probability. Some choices were based on [5]'s article.

V - RESULTS

One evaluation metric was used to evaluate the system. Based upon the Word Error Rate (WER), a BER *-Boxeme Error Rate-* was calculated by adding the number of substitutions, insertions and deletions divided by the number of boxemes in the reference. The better the recognition, the smaller the BER value.

Graphs 4 and 5 give the BER for decoding performances. The "goal" line on horizontal axis represents our objective: obtaining a 10% BER or less, set to guarantee an interesting use of our system by the audience. Graph 4 shows the different performances in decoding for different types of microphones.

Graph 1: BER obtained with monophone acoustic models for the six microphones



The training sets are composed of recordings of the selected microphone to test.

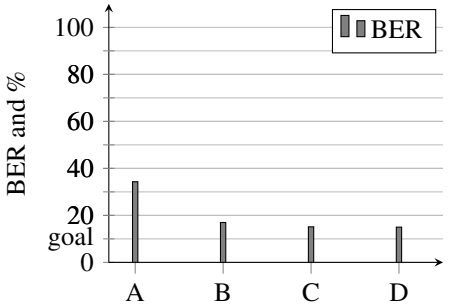
As a result, DPA condenser microphones and Shure SM58 dynamic ones, either placed close or far from the beatboxer's mouth, provide similar performances. Worse recognition rates (with high BER) are found for recordings with encapsulated Shure beta 58 dynamic microphone and Brauner VM1 condenser microphone.

We then decided to vary the silence probability from 0.5 (default) to 0.9 with a step set to 0.1 and tested a 0.99 probability (as it has to be lower than 1). Our best model was achieved with a 0.9 silence probability, which gave a 26,94% BER. When specifying on top a 'pause' phoneme before and after each boxeme in the lexicon, our best results are 16.97% BER. These are obtained with a 0.8 silence probability.

Graph 2 shows the evolution of the results with different parameters : higher silence probability, addition of a pause in the lexicon, 22 MFCC instead of default 13 and 5 HMM instead of default 3. The training set was made with recordings of non-encapsulated microphones. The test set is composed of recordings of the Shure SM58 'sm58p' microphone.

Graph 2: Evolution of BER with different parametrisations

A : default, B: 0.8 silence probability + pause,
C: B + 22 MFCC, D: B + 5 HMM



Our best model is achieved with the C configuration and gives a 15.13% BER. B and D configurations are very close with a 16.97% BER and 16.24% BER respectively (see table 3 for details about the substitutions, insertions, deletions and correct boxeme rates).

In graph 3 and table 3, we observe that every change in parametrisation is beneficial for substitutions, insertions, deletions and correct boxeme rates. The most obvious benefit is for insertions rate which comes up to zero. The correct boxeme rate achieves 85% with the C configurations, meaning that 8.5 boxemes over 10 are well recognised.

	A	**B**	**C**	**D**
Substitutions	19.19%	12.73%	10.70%	12.36%
Insertions	9.41%	0.18%	0.18%	0%
Deletion	5.72%	4.06%	4.24%	3.87%
CBR	75.09%	83.21%	85.06%	83.76%

Table 3: Insertions, substitutions, deletions and Correct Boxeme Rate (CBR) for A B C D configurations
A : default, B: 0.8 silence probability + pause, C: B + 22 MFCC, D: B + 5 HMM

VI - Discussion

As we can see from previous section, the efficacy of the different microphones is quite similar except for the Shure beta 58 and Brauner VM1 microphones which perform worse. We suppose it is because of the way we used them, and independent of the type of microphone. Indeed, the Shure beta 58 microphone is encapsulated and that use affects the performances of the microphone. As for the Brauner VM1 condenser microphone, we can observe it is performing worse than the other condenser microphone in our test (DPA 4060) and suppose it was placed too close from the mouth of the beatboxer. Finally, neither the number of MFCC nor the number of HMM states gives clear improvement. We suppose that having more HMM states could be interesting for complex sounds that are composed of two or more boxemes. This could be analysed in further studies.

VII - Conclusions and perspectives

Our system demonstrates the possibility of using a speech-dedicated recognition system to recognise beatbox sounds. Plus, we also demonstrate the possibility of recognising subtle variations of beatbox sounds as, for example, inhaled and exhaled sounds that are distinguished and not every time mixed up with one another.

So far, our best model was obtained with an increase of the silence probability (0.8 instead of 0.5), the silence phoneme "pause" being added in right and left contexts in the vocabulary and 22 MFCC. The best BER is 15.13%.

We could observe that the type of microphone used for recording does not seem to have any influence on the system. It depends more on their use (encapsulated or not). Putting aside the encapsulated microphone for training gives better results.

As for the different types of production, when mixed, they seem to badly degrade the performance. For now, regarding the substitutions, we can not conclude anything as the system seems to either mix up sounds that are quite similar to the ear or that have quite similar articulation, and sounds that are very different. We suppose that dividing the corpus depending on the sound length and adapting the number of HMM states could improve the system.

Dividing each sound in smaller chunks, as it is done for languages with phonemes or syllables is a perspective. Indeed, as the corpus vocabulary increases, the memory is more and more in demand with word-based speech recognition. Having a boxeme-based model would decrease the number of models needed by the system and enable the treatment of coarticulation. Also, there are still rhythmic sequences and encapsulated sounds recognition to explore. Finally, it would be interesting to see if the difficult recognition of women and children voices in ASR is also a problem in beatbox sounds recognition.

VII - References

[1] A. Contesse and A. Pinchaud. *vocal grammatics*. Web page, www.vocalgrammatics.fr, Last consulted: 2019-08-29, Aug. 2019.

[2] V. Gupta and M. Lennig. Large Vocabulary Isolated Word Recognition. In R. P. Ramachandran and R. J. Mammone, editors, *Modern Methods of Speech Processing*, The Springer International Series in Engineering and Computer Science, pages 213–230. Springer US, Boston, MA, 1995.

[3] K. Hipke, M. Toomim, R. Fiebrink, and J. Fogarty. BeatBox: End-user Interactive Definition and Training of Recognizers for Percussive Vocalizations. pages 121–124, Como, Italy, May 2014. ACM.

[4] A. Kapur, G. Tzanetakis, and M. Benning. Query-by-Beat-Boxing: Music Retrieval For The DJ. Barcelona, Spain, Jan. 2004.

[5] B. Picart, S. Brognaux, and S. Dupont. Analysis and automatic recognition of Human BeatBox sounds: A comparative study. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4255–4259, Brisbane, QLD, Australia, 2015.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi Speech Recognition Toolkit. page 4, Hilton Waikoloa, Big Island, Hawaii, US, 2011.

[7] E. Sinyor, C. McKay, R. Fiebrink, D. McEnnis, and I. Fujinaga. Beatbox classification using ACE. page 4, London, UK, 2005.

[8] V. Tiwari. MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, pages 19–22, 2010.