Human Beatbox Sound Recognition using an Automatic Speech Recognition Toolkit

Solène Evain[a,*], Benjamin Lecouteux[a,**], Didier Schwab[a,*], Adrien Contesse[b,c], Antoine Pinchaud[c] and Nathalie Henrich Bernardoni[d,**]

[a]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[b]ESAD Amiens, De-sign-e Lab, 80080 Amiens, France
[c]http://www.vocalgrammatics.fr/
[d]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

ARTICLE INFO

Keywords:
Human beatbox
automatic speech recognition
Kaldi
isolated sound recognition

ABSTRACT

Human beatboxing is a vocal art making use of speech organs to produce vocal drum sounds and imitate musical instruments. Beatbox sound classification is a current challenge that can be used for automatic database annotation and music-information retrieval. In this study, a large-vocabulary human-beatbox sound recognition system was developed with an adaptation of Kaldi toolbox, a widely-used tool for automatic speech recognition. The corpus consisted of eighty boxemes, which were recorded repeatedly by two beatboxers. The sounds were annotated and transcribed to the system by means of a beatbox specific morphographic writing system (Vocal Grammatics). The recognition-system robustness to recording conditions was assessed on recordings of six different microphones and settings. The decoding part was made with monophone acoustic models trained with a classical HMM-GMM model. A change of acoustic features (MFCC, PLP, Fbank) and a variation of different parameters of the beatbox recognition system were tested : i) the number of HMM states, ii) the number of MFCC, iii) the presence or not of a pause boxeme in right and left contexts in the lexicon and iv) the rate of silence probability. Our best model was obtained with the addition of a pause in left and right contexts of each boxeme in the lexicon, a 0.8 silence probability, 22 MFCC and three states HMM. Boxeme error rate in such configuration was lowered to 13.65%, and 8.6 boxemes over 10 were well recognized. The recording settings did not greatly affect system performance, apart from recording with closed-cup technique.

1. Introduction

Human beatboxing emerged as a vocal practice in the '80s in the Bronx, a borough of New York City. It became part of hip-hop culture. It consists in reproducing all kinds of sounds with one's vocal instrument, especially drum sounds or imitations of musical instruments such as trumpet or electric guitar [13]. Human beatboxers use the same articulators as those of speech. If beatboxing is primarily an outstanding vocal performance, it can also be used as an indexing tool for music information retrieval [5], as a control tool for voice-controlled applications [4] or as the basis of exercises in speech therapy and voice pedagogy.

Very few studies have explored the question of human-beatbox sound classification, whereas its technological and clinical uses grow fast. Good classification rates were obtained with an ACE-based system[1] on a limited range of sound classes, i.e. five main beatbox sounds *bass drum*, *open hi-hat*, *closed hi-hat*, *k-snare* and *p-snare* drums [12, 3]. To the best of our knowledge, automatic recognition of beatbox sounds using a speech recognition system has only been explored by [8]. Their training database consists of isolated beatbox drum sounds (five classes *cymbal*, *hi-hat*, *kick*, *rimshot* and *snare*) and instrumental imitations (8 classes). Performance was poor for imitated sounds (best recognition error rate of 41%), yet good performance was demonstrated for limited beatbox sound classes (best recognition error rate of 9%).

The approach promoted in our study is based on automatic speech recognition systems (ASR). Indeed, most of the work on automatic beatbox recognition is based on classification systems that are independent of the continuous aspect of the signal and/or its rhythmic representation. On the basis of past studies [11, 7], we postulate that human beatbox be considered as a musical language composed of sound units that we call *boxemes* with reference to speech phonemes. Boxemes are co-articulated in beatbox musical phrases. The rhythmic representation can be integrated into the modeling/recognition of beatbox sound production : acoustic components will be based on boxemes and, in the long term, linguistic model will represent the rhythmic aspects. The well-known and widely-used Kaldi ASR toolkit [9] was chosen for this purpose. This toolkit provides state-of-the-art tools in automatic speech recognition.

Can such a commonly-used speech-recognition tool be adapted to build an automatic beatbox-sound recognition system ? This question is addressed in the present study, in an attempt to design an efficient and reliable automatic beatbox sound recognition system that would handle a great number

*Corresponding author
**Principal corresponding author
solene.evain@univ-grenoble-alpes.fr (S. Evain); Benjamin.Lecouteux@imag.fr (B. Lecouteux); Didier.Schwab@imag.fr (D. Schwab); AdrienContesse@gmail.com (A. Contesse); APinchaud@gmail.com (A. Pinchaud); nathalie.henrich@gipsa-lab.fr (N.H. Bernardoni)
ORCID(s):

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

Characteristics

Beatboxers	Adrien (amateur), Andro (professional)
Date	2019
Recording total duration	~206 min
Vocabulary size	80
Number of recorded boxemes per beatboxer	Adrien: 56/80, Andro: 80/80
Writing system used for transcription	Vocal Grammatics
Microphones	5 recorded simultaneously, 1 recorded separately (using closed-cup technique)
Recording parameters	44100 Hz, 16 bits, mono, wav

Microphones

Microphone reference	Label	Type	Distance from the mouth	Usage
Brauner VM1	braun	Condenser	10 cm	with pop filter
DPA 4006	ambia	Condenser ambient	50 cm	
DPA 4060	tie	Condenser	10 cm	tie microphone
Shure SM58	sm58p	Dynamic	10 cm	
Shure SM58	sm58l	Dynamic	15 cm	
Shure beta 58	beta	Dynamic	1 cm	with closed-cup technique

Table 1
Recap chart of the beatbox-VG2019 corpus

of sound classes and enable the recognition of subtle sound variants. The number of sound categories in human beatbox is constantly growing. A system that would take into account more boxemes than the 13 classes of [8]'s study is a current challenge. In addition, this work was made with a view to creating an interactive artistic setup that would provide visual feedbacks during boxeme production. It was intended to be used by professional beatboxers as well as amateurs or beginners. This practical purpose raised the questions of corpus recording condition and robustness to microphone differences. These questions will also be addressed in the present study.

The paper is structured as follows. Section 2.1 presents the training and test databases. The recognition system is presented in Section 2.2. Different experiments are described in Section 2.3 and their results are given in Section 3. Sections 4 and 5 provide a discussion and conclusion to the paper, along with guidelines for future works.

2. Material and Methods

2.1. Corpus, Annotation and Recording Set-up

A dedicated beatbox sound corpus was recorded and named beatbox-VG2019. It is composed of 80 different boxemes, which is a large vocabulary corpus compared to previous corpora. The beatboxer population is predominantly male, so we chose to focus on male voice for the present study and leave gender balance in the recognition system for next step. Two male beatboxers participated in the recordings : a professional beatboxer (fifth author, stage name *Andro*) and an amateur one (fourth author). Only the professional beatboxer recorded samples of all 80 requested boxemes. The amateur beatboxer did not have the ability to perform all boxemes at a good level, so he recorded only 56 boxemes out of 80. The protocol consisted of sequences where boxemes were repeated several times with a pause in between (referred to as isolated sounds in the paper) and additional rhythmic sequences where boxemes were co-articulated in beatbox musical phrases. Only isolated sounds are considered here. Rhythmic sequences will be the target of future studies.

An articulatory-based morphographic writing system developed by the fourth author and called *Vocal Grammatics* [1] was used for annotation. In this system, the glyphs are composed of two pieces of information : the place of articulation (bilabial, glottal, ...), and the manner of articulation (plosive, fricative, ...). Fig. 1 illustrates this writing system in the case of a bilabial plosive with a morphological glyph representing two lips and a cross-shaped glyph symbolising plosion.

The recording session took place in a professional studio. Five microphones were used to record simultaneously the beatboxer's sound production. The microphones differed in terms of specificities (e.g. condenser vs dynamic) and settings. In addition, a separate recording was done with a sixth microphone using a closed-cup technique commonly

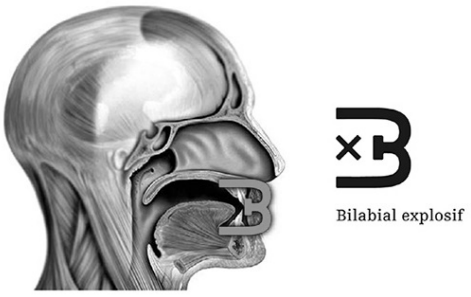


Figure 1: Representation of a bilabial plosive with *Vocal Grammatics* morphographic writing system

found in human-beatbox practice, where one or two hands cover the microphone capsule. Figure 2 shows the setups of all microphones. A DPA 4060 lavalier microphone (tie) was attached to the beatboxer, at 10cm from his mouth. Two identical Shure SM58 microphones were placed respectively at 10cm (SM58p) and 15cm (SM58l) from the mouth. A Brauner VM1 condenser microphone (braun) with a pop filter was placed at same distance than SM58l dynamic microphone. An DPA 4006 ambient condenser microphone (ambia) was placed behind all these microphones, at 50cm away from the beatboxer's face. Finally, a hand helded Shure Beta 58, with the hand leaning on the face, was used for the recording with closed-cup technique.



Figure 2: Placement of all microphones : tie, Braun, SM58 at 10cm and 15 cm, ambiant at 50 cm, and Beta SM58 with closed-cup technique

Table 2.1 is a recap chart which provides full details on the corpus and recording conditions. The different microphones and placements are described. All audio signals were sampled at 44.1 kHz on 16 bits.

The slight acoustic differences between microphonic recordings are illustrated in Figure 3 in the case of a bilabial plosive sound followed by an apico-velar fricative sound (*bilabial_explosif_apico-alvéolaire_fricatif* sound). The recorded acoustic signals differ from one microphone to the other, due to mouth distance, microphone surroundings and transducer proper characteristics. Beta SM58 microphone also differ by the grip technique and the fact that it was not used simultaneously to the other microphones.

2.2. Recognition System

The main goal of our work is to assess whether the automatic recognition of beatbox sounds is possible via a speech-dedicated recognition system. In ASR systems, words are cut into smaller units (e.g. phonemes, syllables) that allow to define a lexicon associating each word with its representation in the form of atomic units. Acoustic models are then trained to recognize these units. Here, we postulate that human beatbox is a musical language that could be similarly structured with distinctive sound units. In support of this assumption, past studies have demonstrated that speech articulators are used to produce beatbox sound units that can be distinguished from each other and that have a specific musical meaning for the beatboxer [11, 7]. These sound units are named boxemes here, in reference to the speech phonemes [7]. Yet in the current implementation, boxemes are altogether the counterpart of speech phonemes and of words.

Two elements are considered distinctly in human beatboxing : acoustic production and linguistic coherence. It lead us to divert a continuous ASR system for the purpose of beatbox sound recognition. Another advantage of continuous ASR system is the ability to work with a lexicon that lists all the words that can be produced. Figure 4 shows the overall operation of an ASR system. It is composed of the following components:

- The acoustic model is trained from sounds associated with their annotations. The acoustic model is trained to recognize basic units (phonemes or boxemes in our case). In our experiments, the acoustic modeling is performed using HMM-GMM models.
- The language model is used to define a probable sequence of events that may occur. A lexicon associates the word and its transcription into phonemes or boxemes. In our case, these words correspond to the different boxemes, considered to be already atomic.
- The role of the decoder is to find the transcription that maximizes the probability of the different models knowing the pronounced sound.

Currently, state-of-the-art implementations for ASR systems are based on Deep Neural Networks (DNN) [2], like ESPnet [15], with either end-to-end or hybrid approaches [10]. End-to-end approaches learn to transcribe a signal directly to its textual transcription. In these systems, DNN learn both acoustic and linguistic representations. Hybrid approaches use Hidden Markov Models (HMM) where transition states are learned via DNN. All these approaches work very well but require quite large amounts of data. Our corpus represents relatively small amounts of data. This led us to use HMM-GMM speech recognition approach. In this approach, acoustic observation likelihoods are computed from a Gaussian Mixture Model (GMM). Due to the assumptions

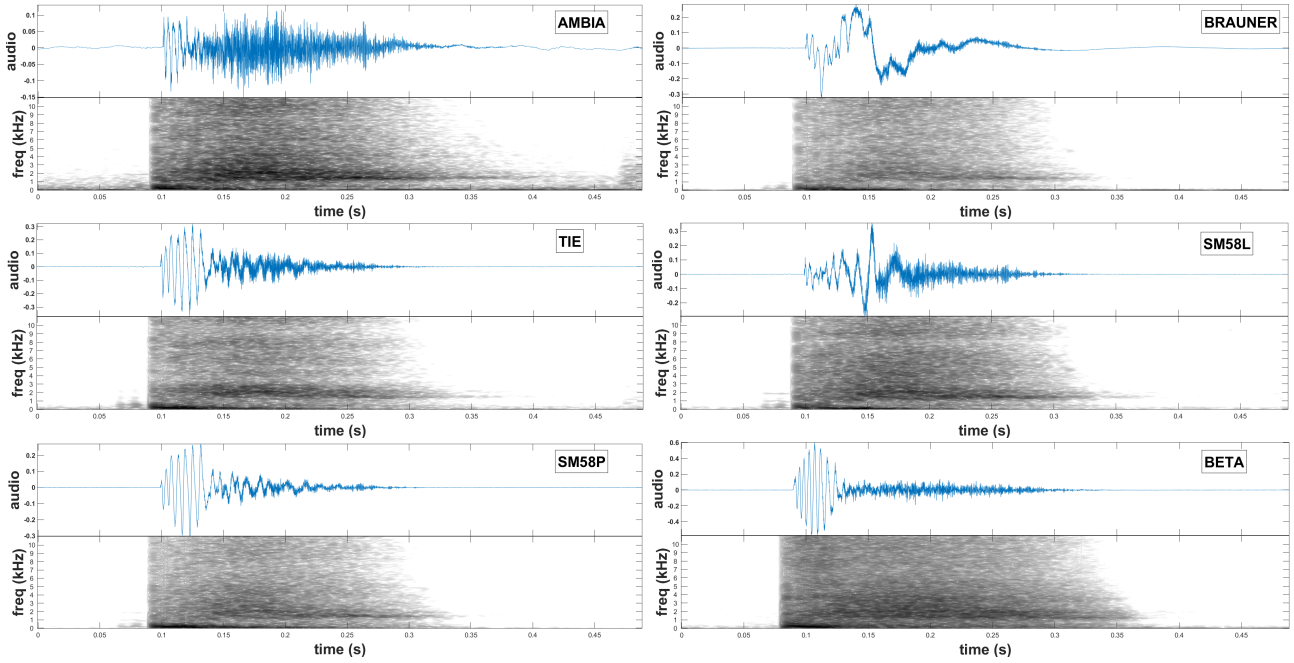


Figure 3: Waveforms and spectrograms of a *bilabial_explosif_apico-alvéolaire_fricatif* 500-ms sound recorded with the six microphones. Audio samples are provided as supplementary material.

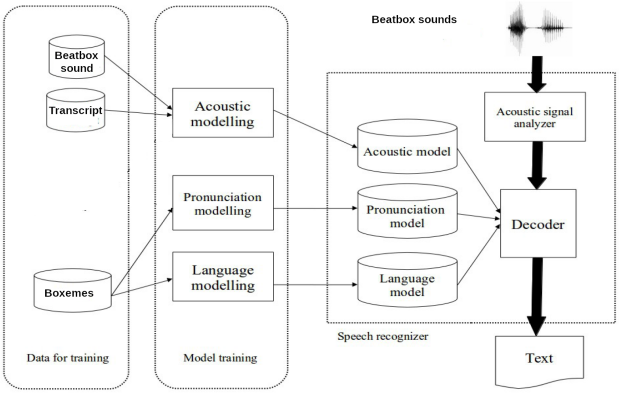


Figure 4: Basics of an automatic speech recognition system, as applied to beatbox sound recognition.

of the HMM-GMM framework, distributions are most accurately modeled for acoustic features that are relatively low-dimensional and somewhat decorrelated. Although this approach is no longer considered to be state-of-the-art in speech, it is at the heart of continuing research efforts, and has been considerably optimized. One advantage of this approach is that it allows acoustic-model estimation with small amounts of data and an easy integration of an expert language model. Another crucial aspect of HMM-based approaches is that they explicitly differentiate the acoustic model from the linguistic one. In our work, this distinction is a necessary requirement.

This first-step work focused on isolated sounds recognition. Co-articulation phenomenon and frontiers between boxeme were discarded, while constraints of noise processing and inter- and intra-beatboxer variability were kept.

The ASR system used to transcript beatbox was trained with the Kaldi speech recognition toolkit [9], widely used in ASR. Several acoustic models were trained on the recorded database :

- different sizes of markov models : the hypothesis is that complex boxemes are difficult to represent with 3-states HMMs that are widely used in automatic speech recognition.
- different resolutions of Mel Frequency Cepstral Coefficient (MFCC) parameters : Features are based on MFCC acoustic features. They are based on human peripheral auditory system [14] and are widely used in ASR.

We focused on monophone-type models. Indeed, the collected corpus presents short pauses between sounds, which suppresses coarticulation effects. A monophone model is an acoustic model that does not include any contextual information about the preceding or following phone. In classic ASR systems, monophones are used as building block for the triphone models, which do make use of contextual information.

In the present work, each boxeme was associated with an entry in the lexicon. In addition, each entry was associated with a HMM. However, as the amounts of data were too small, a speaker adaptation system was not set up.

Another aim of our study was to link Vocal-Grammatics pictographic writing and our beatbox recognition system. Vocal-Grammatics vocabulary is composed of glyphs. The

glyphs were transcripted to text using an analogy with articulatory phonetics. That is how Figure 1 can be described as a "bilabial plosive". Corpus annotation was based on these textual transcriptions. Reversely, the textual transcriptions can be converted back to glyphs as the output of the beatbox sound recognition system.

2.3. Evaluation Methods

The beatbox-VG2019 corpus was split into two parts. Recordings for the five microphones used simultaneously constituted a first subset. A second subset was constituted with acoustic output of beta microphone. Indeed, the latter was recorded separately from the other microphones in a session on its own, and the microphone grip peculiar to closed-cup technique (covering the cup with one or two hands) meant a very different acoustic result for each boxeme (see an illustration in Figure 3).

The performances of the recognition system were evaluated by computing a *boxeme error rate* (BER). Such evaluation metric is inspired from the *word error rate* (WER), main metric applied to ASR evaluation. It is calculated as the total number of error cases (summation of number of substitutions, insertions and deletions) divided by the number of boxemes in the reference. The better the recognition, the lower the BER value.

A second evaluation metric, the *correct boxeme rate* (CBR), was used to rate well-recognized boxemes. It is calculated as the total number of well-recognized boxemes divided by the number of boxemes in the reference.

2.3.1. Recognition robustness and Recording Settings

This main subset with five microphones was used to evaluate the robustness of recognition according to acoustic recording conditions (variability in microphone placement and microphone sensitivity). We aimed to classify the microphones from the less efficient to the most efficient one, and to see whether the use of one of them could really degrade the recognition results. For each microphone, recordings were split into two sets : a train set (with 6 repetitions per boxeme) and a test set (with 7 repetitions per boxeme on average). Both sets are detailed in Table 2.

Then, several configurations of the beatbox recognition system were trained for the purpose of testing different parameters. First, we conducted a comparison on the type of features, namely MFCC, PLP and Fbank. This comparison (see Table 5 in the results part) lead us to select MFCC features for our system. Additional parameters were then varied : i) the number of HMM states, ii) the number of MFCC, iii) the presence or not of a pause boxeme in right and left contexts in the lexicon and iv) the rate of silence probability. A default configuration as proposed by Kaldi system was chosen : 13 MFCC, 3 HMM states, no pause, 0.5 silence probability rate. For varying the number of MFCC, the choice was based on [8], who found their best results for 22 MFCC parameters. The following configurations of the recognition system were tested :

- **Features experiment** : 3 HMM states, 13 MFCC or 13 PLP or 40 FBANK parameters + delta + cmvn , 0.5 silence probability, no pause boxeme in the lexicon ;
- **Configuration A** : 3 HMM states, 13 MFCC parameters + delta + cmvn , 0.5 silence probability, no pause boxeme in the lexicon ;
- **Configuration B** : 3 HMM states, 13 MFCC parameters + delta + cmvn, 0.8 silence probability, addition of a pause boxeme in the lexicon;
- **Configuration C** : same as B configuration, yet with 22 MFCC parameters;
- **Configuration D** : same as B configuration, yet with 5 HMM states.

Raw train set

microphone	number of boxemes	repetitions per boxeme	recording time
ambia	810	6	00:15:18
braun	810	6	00:15:15
tie	804	6	00:15:16
sm58l	810	6	00:15:19
sm58p	810	6	00:15:21

Raw test set

microphone	number of boxemes	avarage repetitions per boxeme	recording time
ambia	952	7	00:19:10
braun	952	7	00:19:08
tie	948	7	00:18:39
sm58l	952	7	00:18:56
sm58p	952	7	00:18:51

Table 2
Details for train and test sets for assessing recognition robustness to acoustic recording conditions (five microphones)

The number of Gaussians was 1000 for 3-HMM configurations and 1500 for 5-HMM configuration.

Table 2 shows that more than 4000 independent sounds were recorded for train and test. However, various problems were encountered : artifacts, noisy sounds, mispronounced sounds,... We therefore manually filtered out problematic sounds to clean the corpus. Only the correctly-pronounced boxemes were kept, as given in Table 3. The choice of using the five microphones for the training was based upon the results of microphone performance testing.

The lexicon of a speech recognition system associates a word to its phonetic transcription : 'word : phonetic_transcription'. In our approach, boxemes take both the place of words and phonemes. The pause boxeme we sometimes added, depending on the system configuration, transforms the lexicon this way : 'boxeme : pause boxeme pause'. This pause

Mic.	Config.	num. of boxemes	rep. per boxeme	recording time
Train set				
ambia, braun, tie, sm58l, sm58p	A,B,C,D	1344	2	00:26:11
Test set				
sm58p	A,B,C,D	542	4	00:09:51

Table 3
Details of train and test sets for decodings with different configurations of the system

is present in the lexicon only. It is not found in the manual transcription of the train and test corpora, so it does not appear in the decoding hypothesis. Therefore, the denominator value of BER is the same for all configurations.

2.3.2. Recognition Evaluation for Closed-cup Technique

The closed-cup technique is widely used by beatboxers on stage. We aimed at checking the recognition efficiency of a training and decoding with such particular sounds which have stronger low frequencies. The details of train and test sets for beta microphone are given in Table 4. First, the A configuration was experimented with no pause in the lexicon. Silence probability, number of HMM states and MFCC parameters were set to Kaldi's default. Then, the configuration in 2.3.1 that gave the best results was chosen to check whether our best configuration for the decoding of the SM58p microphone would also improve the results for beta one.

Microphone	num. of boxemes	repetitions per boxeme	recording time
Train set			
beta	810	6	00:15:07
Test set			
beta	968	7	00:19:14

Table 4
Details of train and test sets for beta microphone

3. Results

3.1. Recognition Robustness for Training with 5-microphones Sub-corpus

Figures 5 to 7 give BER for decoding performances in the case of a training with all recordings of the five microphones (see Table 2). The "goal" line on horizontal axis represents our objective to obtain a 10% BER or less, set to guarantee an interesting use of our system by a beatboxers' audience.

Figure 5 shows the performances in decoding for the five microphones with monophone acoustic models. BER values were found to be high for all microphones, meaning low recognition performances. Similar BER were computed for ambia, tie microphone and sm58 ones, either placed close (sm58p) or far (sm58l) from the beatboxer's mouth : respectively 53.93%, 51.63%, 54.46% and 50.68%. Worse recognition rates with highest BER were found for recordings with braun microphone, a condenser microphone with pop-up filter, leading to a 67.47% BER.

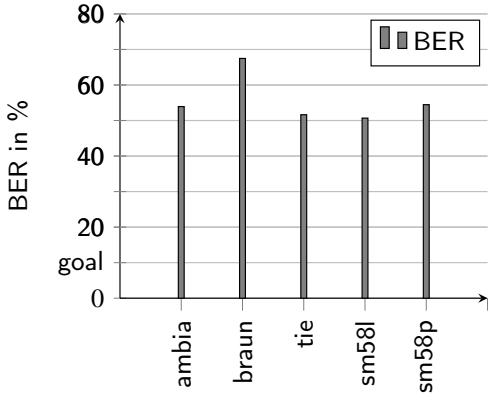


Figure 5: BER obtained with monophone acoustic models for the five microphones

Selecting the sm58p microphone for test set (Table 3), we first tested three different types of acoustic features (MFCC, PLP and Fbank), before exploring different configurations of the beatbox recognition system.

Table 5 shows the results for MFCC, PLP and Fbank features. We observe that MFCC features outperform other ones. PLP features demonstrate a slightly lower recognition quality than MFCC, as already found by [8]. As far as Fbank parameters are concerned, the results are more surprising because they do not seem adapted to the beatbox. We experimented with several numbers of filters (30, 40 and 60) and the best scores are obtained with 40 filters, but the recognition rates remain poor. We assume that there is less covariance between MFCC coefficients than between Fbank outputs : that is important to fit a Gaussian probability density function to the data. Moreover, Fbanks are more efficient when used with neural models, learned from larger amounts of data.

	MFCC	PLP	FBANK
CBR	81.55%	71.96%	59.03%
BER	22.88%	33.58%	42.23%

Table 5
Comparison of MFCC, PLP and Fbank features

Then, the four configurations described in Section 2.3.1 were explored. First, silence probability rate was varied from 0.5 (Kaldi's default) to 0.9 with a step set to 0.1. As the si-

lence probability rate had to be lower than 1, the limit of 0.99 was also tested. Figure 6 shows the evolution of BER results as a function of silence probability, with (red square markers) or without (blue circle markers) adding a pause boxeme in the lexicon (configuration B).

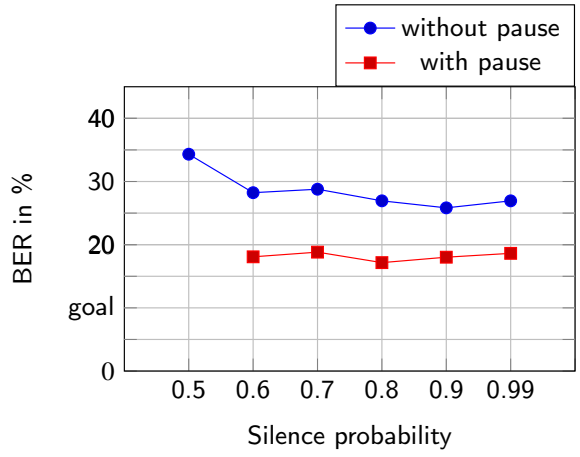


Figure 6: Evolution of BER as a function of silence probability, with or without addition of a pause boxeme in lexicon

With no pause boxeme in lexicon, the greater the silence probability, the lower the BER. A silence probability rate of 0.9 gave the lowest BER of 26,94%.

When adding a pause boxeme to the lexicon (as mentioned in Configurations B, C and D), improvements were obtained with best results for a silence probability rate of 0.8. The BER value was lowered down to 17.16% (configuration B).

Figure 7 shows the recognition results in terms of BER for the four considered configurations. Our best model was achieved for the configuration C, with a BER of 13.65%. Configurations B and D demonstrated slightly worse results with a BER of 17.16% and 15.13% respectively.

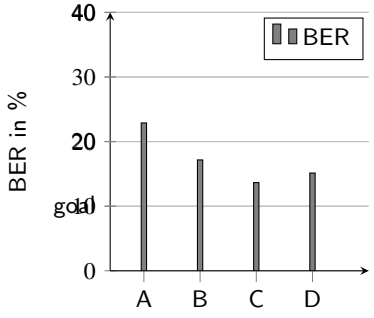


Figure 7: Evolution of the BER for A, B, C and D configurations

A : 3-HMM + mfcc / **B**: prob. silence=0.8 + pause / **C**: B + MFCC=22 / **D**: B + HMM states=5

So as to better understand the system capabilities, Table 6 provides details about substitutions, insertions, deletions and correct boxeme rates. It shows that each change in parametrisation was beneficial for substitutions, insertions, deletions and correct boxeme rates as it lowered them. The most obvious benefit was obtained for the insertion rate which came close to zero for B, C and D configurations. The 22 MFCC parameters in configuration C appeared to be the most beneficial to substitution rate, as errors dropped to 9.78%. Correct boxeme rate reached its highest value of 86.53% with configuration C, meaning that 8.6 boxemes over 10 were well recognized.

	A	B	C	D
Substitutions	14.58%	12.18%	9.78%	11.25%
Insertions	4.43%	0.74%	0.18%	0.37%
Deletion	3.87%	4.24%	3.69%	3.51%
CBR	81.55%	83.58%	86.53%	85.24%
BER	22.88%	17.16%	13.65%	15.13%

Table 6
Percentage of insertions, substitutions and deletions in boxemes' recognition for configurations A, B, C,D, together with corresponding Correct Boxeme Rate (CBR) and Boxeme Error Rate (BER)
A : default, B: 0.8 silence probability + pause, C: B + 22 MFCC, D: B + 5 HMM

3.2. Recognition Robustness for Training with Beta-microphone Sub-corpus

Training and decoding done on beta-recordings sub-corpus with configuration A gave a poor recognition performance with a BER of 70.79% (see Figure 8). We then selected the configuration C that gave us the best results on the five-microphone sub-corpus. A BER of 38.91% was obtained with this configuration, which is 1.8 times lower.

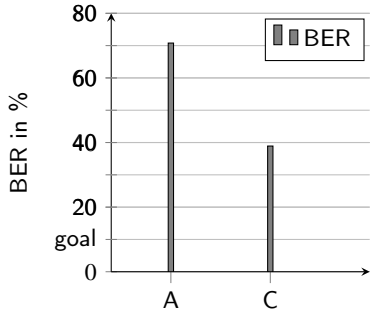


Figure 8: BER values for beta microphone and configuration A and C

Table 7 gives detailed results for insertions, substitutions, deletions, and CBR/BER rates for these two configurations A and C. Similarly to the observed decrease obtained for the number of insertions for the decodings on the SM58p, there was a great improvement of 9.18% reflected on BER. The number of substitutions is also greatly reduced as it lowers from 45.92% to 23.12%, leading to a 22.8 BER points drop. The CBR indicates that 6.6 words over 10 are well recognized, which is still less than for the previous tests with sm58p microphone.

	A	C
Substitutions	45.92%	23.12%
Insertions	14.65%	5.47%
Deletion	10.22%	10.32%
CBR	43.86%	66.56%
BER	70.79%	38.91%

Table 7
Percentage of insertions, substitutions, deletions, Correct Boxeme Rate (CBR) and BER (Boxeme Error Rate) for A and C configurations on beta microphone

4. Discussion

Our results demonstrate the feasibility of using a speech-dedicated recognition system to decode human-beatbox sounds. Our best model achieved a 13.65% BER (Boxeme Error Rate) and 86.53% CBR (Correct Boxeme Rate), meaning that 8.6 boxemes over 10 are well recognized. These results allow an interesting use of our system for artistic purposes.

With parameters set to no pause in the lexicon, 0.5 silence probability, 3 HMM states and 13 MFCC, our system performed a 22.88% BER. We also tested PLP and Fbank features, but they proved to be worse than MFCC. The addition of a pause in the lexicon and the increase of silence probability rate appeared to be very helpful for isolated sounds recognition, giving a 17.16% BER which is much less than the default system. This could be explained by the fact that speech recognition has evolved and that decoding full sentences is now the focus of attention. Being a state-of-the-art toolbox, Kaldi is configured for sentence recognition, and not for isolated words. This is why the addition of a pause in the lexicon and the increase of silence probability helped the system to decode better. By doing so, we intensified the fact that we are working on isolated sounds recognition. The improvement being significant, we chose to make this version of our system a basis for the tests on the number of MFCC and HMM states. As for the other parameters, we observed a slight improvement among our different configurations when the resolution of MFCC is increased. This gave us our best model, with a 13.65% BER. Improvement from default configuration is smaller when the number of HMM states is increased, giving a 15.13% BER. Increasing the MFCC resolution may allow us to capture more finely beatbox-specific signal elements. We thought that increasing the number of HMM states would have a greater impact. Overall, we see that increasing the number of states sometimes degrades the results and improves them in other places. It means that the HMM topology might need to be adapted to beatbox sound types, with 3 states for short sounds and 5 states for more complex ones. Another hypothesis about this result is simply that the amount of data is too small to result in a larger number of states. This will be analyzed in further studies.

Recording settings did not impact much the recognition capabilities of the system. One microphone, Brauner VM1, provided worse results than the other condenser microphone in our test (DPA 4060). As shown in Figure 3, its acoustic signature differed much from the other ones for the same boxeme. Additional pop filter may be a reason for such difference. Indeed, plosive beatbox sounds are produced with strong bursts. The pop filter is supposed to soften them. If we compare the number of substitutions and deletions for plosive boxemes in the results of braun and sm58p testing performances (see table 8), brauner microphone seems to perform slightly worse on those sounds. But if we do the ratio, we see that no conclusion can be drawn. Indeed, the two microphones have the same error rate on those sounds.

	Substitutions	rate	Deletions	rate
brauner	200/361	55%	68/182	37%
sm58p	173/322	53%	26/81	32%

Table 8
Plosive boxeme substitutions and deletions comparison between brauner and sm58p

Finally, the Shure beta 58 gave bad results especially when used with a default configuration of the system. We assume that it is independent of the type of microphone, but that it may be due to how this microphone is held. The closed-cup technique may affect the performances of the microphone. This technique emphasizes low frequencies, which may not be optimal for recognition purpose. Nevertheless, we could observe a much better BER rate when the beta microphone recordings were decoded with the C configuration of our system.

5. Conclusion and Perspectives

Our system demonstrates the possibility of using a speech-dedicated recognition system to recognize human-beatbox sounds. 80 classes were discriminated. To the best of our knowledge, we present the first system able to differentiate such an amount of beatbox sound classes.

So far, our best model was obtained with an increase of the silence probability (0.8 instead of 0.5), a silent boxeme 'pause' being added in right and left contexts in the lexicon, 3 HMM states and 22 MFCC parameters instead of 13. The best obtained BER is 13.65% which appears to be close to our goal of 10%, set for an interesting use of our system by beatboxers during their performances. The CBR indicates that 8.6 boxemes over 10 are well recognized which is quite satisfactory for demonstration purposes. We assume that dividing the corpus depending on boxeme sound duration and adapting the number of HMM states could improve the system, with 3-states HMM applied to short sounds and 5-states HMM to longer ones.

The recording settings, in particular different types of microphone, did not seem to have any influence on the system performance. The difference in efficiency seems to depend more on their use (with or without closed-cup technique, with or without a pop filter) than on the type of microphone or on their distance to the mouth. Poor results ob-

tained with closed-cup recordings, yet commonly used in human beatboxing, call for further studies.

Our corpus is composed of simple and complex sounds. Dividing each sound in smaller chunks, as it is done for languages with phonemes or syllables, is also 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. By doing so, our term *boxeme* would be a real inspiration of the speech *phoneme* and would be distinguishable from the word level which could be composed of many boxemes.

Also, there are still rhythmic sequences recognition to explore. For that purpose, a language model would have to be trained in order to help the system determine the strong probability or not of a row of boxemes. Co-articulation would be studied too. However, we would have to record another corpus with much more beatboxers to have a real representation of the 'beatboxing language'. Only male beatboxers were used for the training set. In order to improve the robustness of the recognition system and its applicability to all genders, female beatboxers should be recorded and added to training corpus.

In terms of more technical perspectives, once we have a larger corpus, we plan to use beatbox recognition methods based on newer technologies, such as deep neural networks. These technologies have made great progress in many areas, but require relatively large amounts of data. One possibility would be to make artificial data augmentation [6] such as in automatic speech recognition. We would also like to explore multimodal approaches, where sensors related to the beatboxer's breathing would complement the automatic beatbox recognition system.

References

[1] Contesse, A., Pinchaud, A., 2019. vocal grammatics. Web page. www.vocalgrammatics.fr. Last consulted: 2019-08-29. URL: http://www.vocalgrammatics.fr/.

[2] Graves, A., Mohamed, A.r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks, in: 2013 IEEE international conference on acoustics, speech and signal processing, IEEE. pp. 6645–6649.

[3] Hazan, A., 2005. Towards automatic transcription of expressive oral percussive performances. In Proceedings of the 10th international conference on Intelligent User Interfaces , 296–298.

[4] Hipke, K., Toomim, M., Fiebrink, R., Fogarty, J., 2014. BeatBox: End-user Interactive Definition and Training of Recognizers for Percussive Vocalizations, ACM, Como, Italy. pp. 121–124. URL: http://dl.acm.org/citation.cfm?doid=2598153.2598189.

[5] Kapur, A., Tzanetakis, G., Benning, M., 2004. Query-by-BeatBoxing: Music Retrieval For The DJ., Barcelona, Spain.

[6] Ko, T., Peddinti, V., Povey, D., Khudanpur, S., 2015. Audio augmentation for speech recognition. URL: https://www.danielpovey.com/files/2015_interspeech_augmentation.pdf.

[7] Paroni, A., Henrich Bernardoni, N., Savariaux, C., Lœvenbruck, H., Calabrese, P., Pellegrini, T., Mouysset, S., Gerber, S., 2021. Vocal drum sounds in human beatboxing: An acoustic and articulatory exploration using electromagnetic articulography. The Journal of the Acoustical Society of America 149, 191–206. URL: https://doi.org/10.1121/10.0002921, doi:10.1121/10.0002921, arXiv:https://doi.org/10.1121/10.0002921.

[8] Picart, B., Brognaux, S., Dupont, S., 2015. Analysis and automatic recognition of Human BeatBox sounds: A comparative study, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, Australia. pp. 4255–4259. doi:10.1109/ICASSP.2015.7178773.

[9] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi speech recognition toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[10] Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., Khudanpur, S., 2016. Purely sequence-trained neural networks for ASR based on lattice-free MMI.

[11] Proctor, M., Bresch, E., Byrd, D., Nayak, K., Narayanan, S., 2013. Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study. The Journal of the Acoustical Society of America 133, 1043–1054. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3574116/, doi:10.1121/1.4773865.

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

[13] Stowell, D., Plumbley, M.D., 2008. Characteristics of the beatboxing vocal style. Tech. Rep., Centre for Digital Music Dep. of Electronic Engineering, Univ. of London .

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

[15] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N.E.Y., Heymann, J., Wiesner, M., Chen, N., et al., 2018. Espnet: End-to-end speech processing toolkit. URL: http://dx.doi.org/10.21437/Interspeech.2018-1456.