A Compact Spectrum-Assisted Human Beatboxing Reinforcement Learning Tool On Smartphone

Simon Lui
Singapore University of Technology and Design
20 Dover Drive, Singapore, 138682
simon_lui@sutd.edu.sg

ABSTRACT
Music is expressive and hard to be described by words. Learning music is therefore not a straightforward task especially for vocal music such as human beatboxing. People usually learn beatboxing in the traditional way of imitating audio sample without steps and instructions. Spectrogram contains a lot of information about audio, but it is too complicated to be understood in real-time. Reinforcement learning is a psychological method, which makes use of reward and/or punishment as stimulus to train the decision-making process of human. We propose a novel music learning approach based on the reinforcement learning method, which makes use of compact and easy-to-read spectrum information as visual clue to assist human beatboxing learning on smartphone. Experimental result shows that the visual information is easy to understand in real-time, which improves the effectiveness of beatboxing self-learning.

Keywords
Audio analysis, music learning tool, reinforcement learning, smartphone app, audio information retrieval.

INTRODUCTION
Human beatboxing is kind of vocal music art, which makes use of lip, tongue and throat to produce percussion and sound effects. It is not easy to learn human beatboxing. The most common way to learn this is to listen to beatboxing music sample and imitate it. There does not exist well-structured and effective learning tool. The main reason is beatboxing require complicated physical technique, which is hard to describe by words and steps. Self-learning by listening to music sample is still the best way to understand beatboxing. In this work, we are looking for reference and clue to assist this self-learning process. Spectrogram contains rich information of voice. For example, Zue [1] demonstrated how to read English words directly by visualizing spectrogram. Spectrogram should be useful for assisting beatboxing learning by providing visual information of the voice. However, the information in a spectrogram is hard to read and understand. We propose to redesign and simplify the spectrogram, which act as a visual clue for user to read in real-time to effectively support the beatboxing self-learning process.

1. Background and Previous Work
One of the most popular beatboxing self-learning app on the Android smartphone platform is *Learn to Beatbox* [2]. It is a traditional unidirectional education tool. It links users to various demo videos without any interaction mechanism to receive input from user during the learning. It is simply a collection of well-organized tutorial videos and users are not guided to improve themselves.

On the Mac OSX platform, *Garageband* provides the *learn to play* music tutorial [3]. User is guided by the video demo. They can interact with it by following the input sequence of either a set of guitar chord positions or piano fingerings. However there is no sound analysis component. It only teaches user on button control and fingering. There are many other music learning apps on the iOS smartphone platform such as *Piano man* [4], G*uitar lab* [5], *Karajan* [6], etc. However all of them are similar to *Garageband* in that they only guide users to follow a certain fingering sequences.

There is no vocal music educational tool that analyses the sound and provides feedback for the users to try out and gradually improve themselves. Furthermore, beatboxing is more than pitch control, requiring fairly complicated mastering of voice through lip, throat and mouth shape [7]. To learn beatboxing effectively, we need some guide to help user to view and adjust their voice.

2. Design Principle and Implementation
We want to design a user-friendly self-learning assisting tool for human beatboxing. There are a lot of features that affect the sound of beatboxing. Kapur identified beatboxing sound with several features such as Root Mean Square (RMS) energy, Mel-frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficient (LPC) [8]. Lemaitre et al. used several audio features to identify vocal imitation, including fundamental frequency, spectral centroid, and energy envelope [9]. These features can be classified into two main classes: the time domain energy features class, and the frequency domain spectral features class. There are other features that can affect the beatboxing sound. Since we aim to design a visual clue for beatboxer to follow, we only pick those features that a beatboxer can directly relate to in controlling their own physical action.

In this work, we pick three features to observe: volume, frequency and mouth shape. Beatboxing involves blowing air from the lung, which controls the volume. Then the air goes through the throat, which controls the frequency. Finally the air comes out through the lips and being spoken out, with the lips and mouth shape controlling the syllable. In order to visualize these three features, we use the root mean square value of the signal amplitude to represent the volume, a simplified spectrogram to represent the frequency, and the formants to represent the mouth shape. It is generally agreed that the root mean square value of signal amplitude represents the volume. It is also well proven that the first two formants are enough to identify the vowel in speech [10]. We use a highly simplified spectrogram to represent the frequency, since such a simplified representation is sufficient for the purpose of a quick guide for the user to visualize the position of syllables and to roughly identify the pitch trend. The three features are displayed in sync with the audio sample playback. The user will sing along and interact with the app, and the app will help the user visualize mistakes and hence correct them through repeated trials.

This work is designed for smartphone as its ubiquitous use by general public today and its simple touch and shake interaction modality suits well for our envisaged non-professional beatboxing training context. We first implement it on Apple devices running iOS 6. We will implement it on other smartphone platforms such as Android and Windows Phone in

the future. A recent research shows that electronic device users are device sensitive when working on different tasks. They take the advantage of the best feature of each kind of device [11]. For example, 90% users send email with computer because of its large screen and keyboard. On the other hand, 73% of navigation activities are performed on the smartphone because of its mobility. The learning process of beatboxing only involves singing and listening without requiring typing, thus a modern smartphone, as opposed to a desktop PC with keyboard, fits for learning beatboxing. All iOS smartphone contains screen with multi-touch control, microphone and speaker, which is all in one in functionality and very convenient for the user. Smartphone app also benefits from its mobility. User can use the app everywhere. However smartphone app needs to run efficiently with limited processing power and limited resources. Also, we need to display a lot of information on the small screen of a smartphone, requiring an extra consideration in designing the user interface. The displayed information needs to be compact and easy to read in real-time.

Male and female beatboxers are different in vocal frequency range. For simplicity we first work on male beatboxer. Without any loss of generality this work will be further developed for female beatboxer in a very similar way.

Figure 1 shows the feature representation of the app. The timeline moves from the right to the left. The volume contour versus time is the main trend for the user to follow. The spectrogram is simplified into only a few frequency portions, and normalized by the volume contour. Only a few formants are selected and they are also normalized by the volume contour. Since volume contour is easy to trace and follow, the benefit of the normalization is that the information can be packed into one single "symbol like" object. User can read both volume and frequency information at one glance. Summation of all frequency components' amplitude at a certain time should equal to the overall amplitude in the time domain. This nice combination preserves the ratio between different frequency components. Another reason for the normalization is that the visual representation is just an aid for learning. User only needs to know the relative frequency ratio and the relative position of the formant. For example, when a beatboxer is guided to tune down his first formant frequency a little bit, it would not be very meaningful to ask him to, say, tune it down exactly by 10.6Hz. Instead, the app should just ask him to tune it down "by a small portion." Then the beatboxer should do it according to his experience and control his muscle, listen to the outcome, and further adjust it according to the next feedback stimulus.

User follows the clue and sings beatboxing note along the timeline. Then the app will rate the user's performance by comparing it with the preset features. This is the reinforcement learning part. *Reinforcement* is the delivery of stimulus resulting from certain action. Human tends to repeat the same behavior according to positive result that they received before [12], and this is how reinforcement learning works: the user is guided by the system through a continuous learning process to approach to the correct answer; the system then provides feedback for the user that encourages improvement; in turn the user will adjust his performance according to the system feedback. Reinforcement learning is especially suitable when a system only knows the correct answer, where the exact steps toward success are missing [13]. There are positive and negative reinforcement, which depends on the kind of stimulus used. Not all stimuli triggered by a certain action can be regarded as reinforcement. For example, a robber may feels happy (positive) for the money he stole but yet feeling guilty (negative). Hence he might not want to steal again in the future, and he is not *reinforced*. There are various kinds of feedback messages to choose. We will look for the best feedback messages in our experiment.

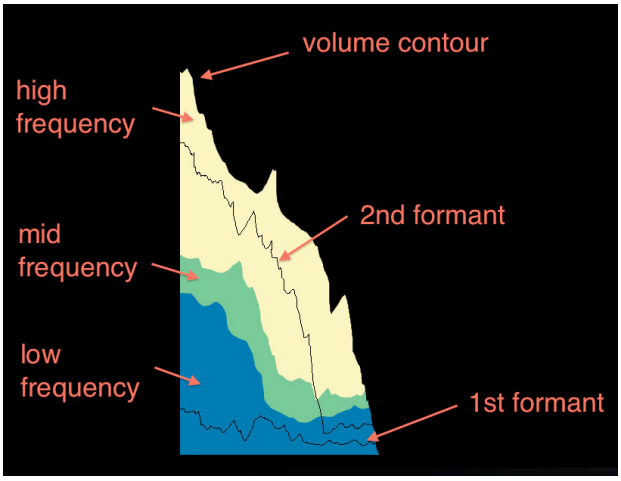

Figure 1. Features representation.

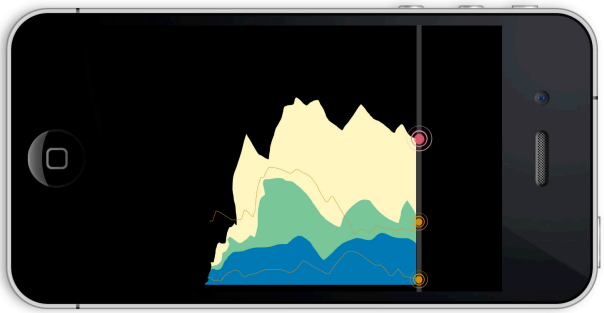

Figure 2. Free mode.

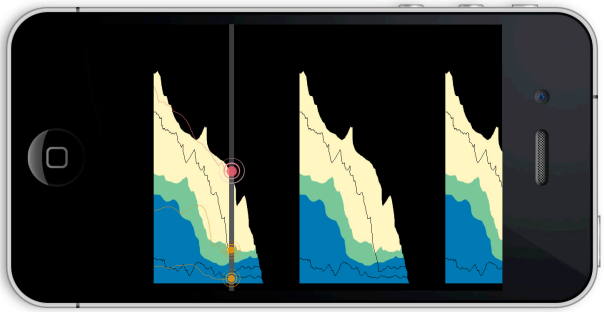

Figure 3. Note-Training mode.

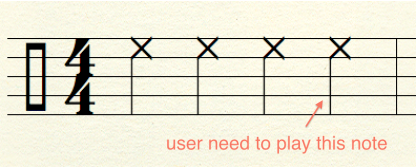

Figure 4. The music used in the note-training mode.

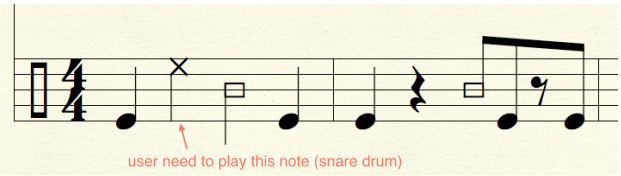

Figure 5. The music used in the phrase-training mode.

We offered three learning modes in the app. In the free mode as shown in Figure 1Figure 2, user simply plays beatboxing notes and observes the visual clue. This is for the user to get familiar with the app interface and the visual clue responses. User will be trained to associate their voice with the consequent visual clue that appears; hence they learn to change the formant position by adjusting their voice. In the note-training mode as shown in Figure 3, a single note's spectrum is displayed in loop, with three voiced note and one unvoiced note for the user to fill out with his/her voice. The top red dot represents user's current volume level. The bottom two orange dots represent user's current formant position. This is for the user to focus on training with one particular kind of beatboxing note. The phrase-training mode is similar to the note-training mode, where the beatboxing drum loop is played recursively. Only one target note is unvoiced for the user to fill it out. Figure 4 and Figure 5 shows the training music used in the note-training mode and phrase-training mode respectively. This mode aims at teaching the user to get the feeling of how a particular beatboxing note sounds in a music phrase. The three training modes provide the user with a complete experience of learning beatboxing in the fundamental, note and phrase perspectives.

3. Experiment and Discussion

We performed several tests to gather user feedback in order to improve the design of the app. We invited a focus group (FG) to perform the tests, consisting of 36 adult males divided into 3 sub-groups: 12 have musical training with beatboxing experience (S1), 12 have musical training but have no beatboxing experience (S2), and 12 non-musicians (S3). All members are within age of 18-45 and evenly distributed in the 3 sub-groups. Six iPod Touch (5th generation) were used. These iOS devices were installed with iOS 6.0, Apple 1GHz dual core A5 CPU, with speaker frequency response of 20Hz to 20000Hz. FG members were invited to perform the test each in a separate and quiet room. They were required to use the app with mouth-to-microphone distance of 15-18 cm.

In the first test we investigated on the effectiveness of learning beatboxing using the 3 features. Each member had 20 seconds to try out the free mode, then 40 seconds for the note-training mode, followed by 40 seconds for the phrase-training mode. Some members used the original app and some members used the app with several visual clues missing. After the trial, their performances were compared and recorded. We used the root mean square value to measure the volume difference, and the harmonic-amplitude error metric to measure the formant and spectrogram difference:

$$\frac{1}{N}\sum_{n=1}^{N}\sqrt[a]{\sum_{k=1}^{K}|A_k(t_n)-A_k'''(t_n)|^a} \quad (1)$$

with K harmonics, amplitude A, and an arbitrary exponent a equal to 0.6 which is proven to be optimal by Horner [14].

The result of the test is as shown in Table 1. It is found that all the three sub-groups, S1 (musicians with beatboxing experience), S2 (musicians without beatboxing experience) and S3 (non-musicians) members performed better with the visual clue, implying the effectiveness of this tool in supporting the beatboxing learning. The volume contour greatly improved S3's performance, while the formant was not very helpful for S3. We believe that the volume is the most understandable information for non-musician, while formant is very difficult for them to understand. On the other hand, all the three features helped S1 and S2 to perform better. The volume contour helped S1 and S2 to a small degree, while the formant and spectrogram helped them a lot. We believe that human with music training can already judge the volume contour by ear without looking at the visual clue. However, the formant and spectrogram information may not be easily identifiable by musician's ear in real-time. Musician can quickly understand the frequency domain information in the visual clue, which is helpful for them to adjust their beatboxing sound quality. There is no big difference among the standard deviation data; hence members in the same sub-group performed similarly under different settings.

Table 1. Test 1 results, effectiveness of the features.

			Error Metric Average			Standard Deviation		
Volume	Formant	Spectrogram	S1	S2	S3	S1	S2	S3
o	o	o	0.311	0.302	0.671	0.082	0.092	0.224
o	x	x	0.489	0.502	0.713	0.110	0.079	0.230
x	o	x	0.304	0.387	0.852	0.092	0.110	0.262
o	x	o	0.342	0.402	0.763	0.093	0.132	0.213
o	o	x	0.354	0.335	0.721	0.121	0.890	0.178
x	x	x	0.552	0.567	0.913	0.135	0.164	0.306

In the second to fifth tests, members were invited to work on the note-training mode for 1 minute. Before each test, they could optionally practice with the free mode for any duration.

In the second test we investigated the effectiveness of user interface representations. We compared different number of formants and spectral lines. Result of the test is as shown in Table 2. The result of test 1 using the original app is also included in the second line of Table 2. Since music concept and principle are universal among different instrument players, it is not surprising to find that S1 and S2 performed similarly, and both S1/S2 performed better than S3. It is found that using the average formant value did not improve the performance compared with using the real value. We believe that the real value contour of formant contains useful information that is lost in the average value. For example, according to FG members' feedback, the first formant's real value contour is related to the force that they should apply in the throat. We found that the members performed better with less spectral line and less formant. FG members commented that it is difficult to read more than 2 spectral lines and 2 formants at the same time. They seldom read the 3rd or more spectral lines or formants. These extra lines usually complicate the information and the members eventually get lost. There is also no big difference among standard deviation of different settings.

Table 2. Test 2 results, effectiveness of choosing different number of features.

		Error Metric Average			Standard Deviation		
Formant	Spectrogram	S1	S2	S3	S1	S2	S3
2 formants, real value	2 spectral lines	0.281	0.312	0.432	0.071	0.072	0.193
3 formants, real value	2 spectral lines	0.311	0.302	0.671	0.082	0.092	0.224
2 formants, average value	2 spectral lines	0.339	0.352	0.421	0.088	0.093	0.201
3 formants, average value	2 spectral lines	0.322	0.338	0.652	0.073	0.083	0.214
2 formants, real value	3 spectral lines	0.293	0.336	0.343	0.079	0.092	0.178
2 formants, real value	4 spectral lines	0.421	0.462	0.675	0.092	0.141	0.253

In the third test, we compared different reinforcement representations. We compared four types of reinforcement feedback messages: the original formant and volume line; the line with deviated portion shaded; the deviation percentage; and just a right (error metric <= 0.2) or wrong (error metric > 0.2) message. We chose an error metric of 0.2 as the threshold, since Horner has proven that an error metric of 0.1 is almost indistinguishable by ear, where an error metric of 0.3 starts to be distinguishable [14]. Hence we take a loose middle point to allow room for acceptable error in beatboxing. Result of the test is as shown in Table 3. It is found that S1's and S2's responses are very different from S3's. S3 performed better with just a right or wrong feedback, while S1 and S2 performed better with the deviated portion shaded. We believe that S3 preferred the simplest information that they can understand, while S1 and S2 were looking for more detailed feedback message. According to S1 and S2 member's feedback, the shaded portion clearly

instructed them to tune the formant either up or down, and either by a large or small scale.

Table 3. Test 3 results, effectiveness of various reinforcement representation

	Error Metric					
	Average			Standard Deviation		
Feedback message	S1	S2	S3	S1	S2	S3
lines	0.311	0.302	0.671	0.082	0.092	0.224
lines with deviated portion shaded	0.267	0.288	0.723	0.093	0.132	0.241
deviated percentage	0.341	0.361	0.692	0.087	0.088	0.302
just right or wrong	0.332	0.329	0.505	0.078	0.073	0.154

In the fourth test, we compared different reinforcement schedule. We chose the deviated percentage of the features and right or wrong feedback as reinforcement message, since the deviated portion cannot be used in scheduling and must be presented in real-time. We displayed the message on the top of the smartphone screen. We worked on the variable ratio (per note), variable time (per time), fixed ratio, fixed time, and we tried different rate for the fixed schedule. Result of the test is as shown in Table 4. It is found that there is no significant advantage for S1 and S2, which match the result in test 3. The fixed ratio schedule improved the performance of S3. According the S3 member's comment, only the fixed ratio schedule gave feedback in regular basis under their expectation. It shows that S3 user count the time with note interval instead of seconds during the test.

Table 4. Test 4 results, reinforcement schedule.

		Error Metric					
		Average			Standard Deviation		
Feedback message		S1	S2	S3	S1	S2	S3
Original, no feedback		0.311	0.302	0.671	0.082	0.092	0.224
Variable ratio	1-4 notes	0.323	0.313	0.743	0.072	0.083	0.243
Variable time	3-9 seconds	0.321	0.353	0.656	0.073	0.063	0.220
Fixed ratio	1 note	0.367	0.361	0.587	0.082	0.072	0.167
Fixed ratio	2 notes	0.312	0.322	0.532	0.094	0.083	0.198
Fixed ratio	3 notes	0.298	0.317	0.435	0.082	0.072	0.132
Fixed ratio	4 notes	0.327	0.296	0.532	0.121	0.102	0.172
Fixed time	3 seconds	0.302	0.283	0.621	0.093	0.132	0.198
Fixed time	5 seconds	0.265	0.314	0.602	0.078	0.129	0.232
Fixed time	7 seconds	0.346	0.325	0.732	0.110	0.111	0.266
Fixed time	9 seconds	0.314	0.329	0.656	0.104	0.098	0.298

In the last test, we compared between positive and negative reinforcement. For the positive reinforcement, we emphasized the correct answer (error metric <= 0.2) with large, bold and highlighted text. We do the reverse for negative reinforcement. Result of the test is as shown in Table 5. The result shows that there is almost no difference between positive and negative reinforcement learning. We can also conclude that there is a little advantage in the positive reinforcement learning approach. We believe that human prefers encouragement instead of punishment in art learning.

Table 5. Test 5 results, reinforcement learning.

	Error Metric					
	Average			Standard Deviation		
Feedback message	S1	S2	S3	S1	S2	S3
original, no emphasize	0.311	0.302	0.671	0.082	0.092	0.224
positive reinforcement	0.276	0.293	0.643	0.102	0.088	0.263
negative reinforcement	0.302	0.332	0.667	0.078	0.121	0.232

Here are some further discussions about the five tests. Visual reinforcement is helpful for user to improve the beatboxing learning process as shown in test 1. The visual clue reduced the error rates of all kinds of participants. Simplified reinforcement performs better than detailed reinforcement in general as shown in test 2. Different people have different need for the degree of information simplification, which is demonstrated by the result of S1, S2 and S3 in the five tests. Non-musician prefers very simple message such as yes/no or right/wrong, which are scheduled regularly in message interval. Trained musician prefers real-time and simplified information extracted from the frequency domain, which provide them with extra information that is not easy to be identified by ear in real-time.

4. Future Work

This work is an early prototype and hence the UI is just functional and experimental. We will revise the UI in terms of usability and aesthetics and publish on the app store. The next work will be designed in a structured way that takes account of the user experience. This will be done by gathering feedback from user reviews as inspired by a similar work of Erkut [15]. We will work with psychologists to further investigate the ways to integrate the reinforcement learning aspect into the user interaction. We also expect to further develop this app into other music learning tool or speech learning tool, especially for the disabled people to learn and understand speech and music. We will work on female beatboxing in the future version. We will combine the 3 features more closely so that they can be more unified and hence easier to read in real-time. The new UI design might be in 3D.

5. ACKNOWLEDGMENTS

This work is supported by the SUTD-MIT International design center (IDC) Research Grant (IDG31200107 / IDD11200105 / IDD61200103). Thanks to Dr. Hyowon Lee for his advice and comment on this work.

6. REFERENCES

[1] Zue, V. 1985. "The Use of Speech Knowledge in Speech Recognition," Special issue on Man-Machine Communication, Proc. IEEE, vol. 73, pp. 1602-1615

[2] https://play.google.com/store/apps/details?id=com.v1_4.B97FD5259156F3EFD3A7996B.com

[3] http://www.apple.com/ilife/garageband/what-is.html

[4] http://www.yudo.jp/

[5] http://truefire.com/apps/

[6] http://www.karajan-eartrainer.com/en/

[7] Stowell, D., and Plumbley, M. D. 2008. "Characteristics of the beatboxing vocal style" Tech. Rep. C4DM-TR-08-01, Dept. of Electronic Engineering, Queen Mary, University of London.

[8] Kapur, A., Benning, M. and Tzanetakis, G. 2004. "Query by beatboxing: Music information retrieval for the dj," in Proc. ISMIR, pp. 170–178.

[9] Lemaitre, G., Dessein, A., Susini, P. and Aura, K. 2011. "Vocal imitations and the identification of sound events". Ecological Psychology 23, 4. pp. 267-307

[10] Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. 1952. An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. Word 8, 195–210.

[11] http://www.businessnewsdaily.com/3691-smartphone-vs-computer-use.html

[12] Schacter, D. 2012. "Psychology". Worth Publications.

[13] Sutton, R. S. and Barto, A. G. 1998. "Reinforcement learning: An introduction." Cambridge, MA: MIT Press.

[14] Horner, A., Beauchamp, J., and So, R. 2006. "A Search for Best Error Metrics to Predict Discrimination of Original and Spectrally Altered Musical Instrument Sounds," Journal of the Audio Engineering Society, 54(3), pp. 14.

[15] Erkut, C., Jylhä, A., and Disçioğlu, R. 2011. "A structured design and evaluation model with application to rhythmic interaction displays". In Proc. NIME '11, pages 477-480.