INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Comparison of Basic Beatboxing Articulations Between Expert and Novice Artists using Real-Time Magnetic Resonance Imaging

Nimisha Patil, Timothy Greer, Reed Blaylock, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory
University of Southern California, USA
nimishhp@usc.edu, timothdg@usc.edu, reed.blaylock@gmail.com, shri@sipi.usc.edu

Abstract

Real-time Magnetic Resonance Imaging (rtMRI) was used to examine mechanisms of sound production in five beatboxers. rtMRI was found to be an effective tool with which to study the articulatory dynamics of this form of human vocal production; it provides a dynamic view of the entire midsagittal vocal tract and at a frame rate (83 fps) sufficient to observe the movement and coordination of critical articulators. The artists' repertoires included percussion elements generated using a wide range of articulatory and airstream mechanisms. Analysis of three common beatboxing sounds resulted in the finding that advanced beatboxers produce stronger ejectives and have greater control over different airstreams than novice beatboxers, to enhance the quality of their sounds. No difference in production mechanisms between males and females was observed. These data offer insights into the ways in which articulators can be trained and used to achieve specific acoustic goals.

Index Terms: speech recognition, speech production, paralinguistics, singing, MRI, beatboxing

1. Introduction

Beatboxing is a musical art form in which artists aim to emulate physical instruments using vocal percussion. Various precursors to modern day beatboxing have been prevalent throughout history, dating centuries back to the tabla bols used in North Indian music, where these vocal bols are used to imitate the tabla drums. Several other forms of vocal percussion – from the intentionally raspy vocal quality used in African spiritual music to the scatting and bass hums in jazz and blues music – have evolved since then. The genesis of early mainstream beatboxing, or "old school" beatboxing, occurred primarily in New York, where the rising hip-hop culture of the 1970's and 80's gave a platform for artists to emulate drum machines, or "beat boxes," as accompaniment to singers and rappers [1].

Since then, beatboxing has become more prevalent in mainstream culture, a cappella singing groups, and even as a stand-alone art form. In the past few years, beatboxers such as Alem and NaPoM have pushed the bounds of this art even more, utilizing intricate patterns while focusing on speed, technicality, musicality, and bass-heavy sounds to develop a style that is now known as "new school" beatboxing.

Though beatboxing has been acknowledged as an art, the science behind its technique has yet to be studied and documented extensively. There are tutorials on how to make several beatboxing sounds available on websites such as YouTube. However, these tutorials lack proper documentation and also fail to provide precise phonetic instruction on how to create the sounds. A few preliminary research studies on beatboxing have been conducted.

TyTe and Splinter [2] developed Standard Beatboxing Notation (SBN) as a way of writing common beatboxing sounds and patterns. Stowell and Plumbley [3] examined different vocal techniques used in beatboxing and concluded that beatboxers use non-syllabic patterns, inhaled sounds, growls, falsetto and trills to generate diverse sounds. Lederer [4] provided a detailed acoustical study of certain beatboxing sounds and concluded that the accuracy of the imitation depends on the nature of target sound and whether or not it is found commonly in language [4]. Torcy et al. [5] examined the laryngopharyngeal behavior during certain beatboxing sounds using fiberoptic imaging to conclude that beatboxers move parts of their laryngopharynx separately. Only one other study by Proctor et al. [6] used rtMRI, and examined 17 sounds produced by one beatboxer to conclude that these sounds were similar to those found in human speech [6].

Though these studies have laid the initial groundwork for examining the production techniques behind beatboxing, they are limited in scope as they only examine the sounds of one beatboxer. The study presented in this paper differs from previous studies by examining and comparing data from five individual subjects of different skill levels and genders.

In this paper, we compare the vocal production mechanisms of three fundamental beatboxing sounds: the kick drum, the closed hi-hat, and the PF snare. These sounds were named according to common beatboxing terms, as these are standard in most beatboxers' repertoires [1, 2]. We show that there is no difference in production mechanisms between males and females. We also show that advanced beatboxers use stronger ejectives and have greater stylistic control over different airstream usage than novice beatboxers.

Multimedia files of scans discussed in this paper can be found at http://sail.usc.edu/span/beatboxing2017/index.html

2. Methods

An MRI machine was used to acquire dynamic images of human articulators of interest along the entire midsagittal vocal tract, while recording the beatboxing sounds and spoken passages of two male beatboxers and three female beatboxers. All subjects reported English as their native language. The only multilingual subject was the intermediate female (first author of this paper), who also spoke Marathi and Hindi fluently. The skill level of each beatboxer – advanced, intermediate and novice – was determined acoustically by the level of artistic control they exhibited over beatboxing sounds and patterns, along with the perceived difficulty of the sounds in their repertoires. Subject information is summarized in Table 1:

http://dx.doi.org/10.21437/Interspeech.2017-1190

Table 1: *Imaged Subjects*

Initials	Skill Level	Gender
AM	Advanced	Male
AF	Advanced	Female
IF	Intermediate	Female
NM	Novice	Male
NF	Novice	Female

The participants were asked to produce all the percussion effects in their repertoire and perform some beatboxing sequences in short intervals while laying supine in an MRI scanner. Though beatboxing is usually performed in an upright position, the imaged difference between articulators in upright versus supine elicitations is typically minimal [7].

For each sound, the elicitation was repeated at least three times in a single MRI recording and subsequently used in a sample beat pattern. Some speech passages were also recorded, and a full set of the subject's American English vowels was elicited using the [h_d] corpus [6]. The subjects were paid for participation in the experiment. The study presented in this paper draws from a subset of this data.

An rtMRI protocol developed to study dynamic vocal tract shaping during speech was used to acquire the data. The scan parameters included: a gradient echo pulse sequence ($T_R = 6.004$ ms); a conventional GE Signa 1.5 T scanner ($G_{max} = 40$ mT/m, $S_{max} = 150$ mT/m/ms); an 8-channel upper-airway custom coil; scan slice thickness of 6 mm over a 200 mm x 200 mm field-of-view; an image resolution of 84 x 84 pixels (2.4 mm x 2.4 mm) [8]. The scan plane was manually aligned with the midsagittal plane of the subject's head. The images were retrospectively reconstructed to a temporal resolution of 12 ms (2 spirals per frame, 83 frames per second) using a temporal finite difference constrained reconstruction algorithm [8, 9] and a recent open-source library [10].

Audio was recorded while the subjects were imaged as well. A custom fiber-optic microphone system along with a sampling frequency of 20 kHz was used to acquire the audio recordings. These recordings were noise-canceled and reintegrated with the reconstructed MRI video [11].

This method provides data that allows for dynamic visualization of the subjects' midsagittal vocal tracts along with synchronized audio. The synchronized audio recordings are particularly useful for studying beatboxing sounds, as beatboxing is an acoustical art form. Qualitative description of the sound production outside the MRI machine was used to supplement articulatory observations.

The main articulators (notably the lips, tongue, velum, and glottis) were observed over several elicitations of the sound to examine production mechanisms. Because the scan plane was in the midsaggital plane of the glottis, it was possible to observe glottal abduction and adduction, as well as larynx raising and lowering. Glottal and velar closures were observed to determine the airstream mechanisms used by the subjects while producing the sounds.

3. Results

Three basic beatboxing sounds were compared across each of the five subjects: the kick drum, the closed hi-hat, and the PF snare (Table 2). These three sounds could be described by International Phonetic Alphabet (IPA) notation, likely due to their derivation from sounds common to language [4].

Table 2: *Isolated Sounds Studied in this Paper*

Sound	IPA	SBN
Kick Drum	/pf'/	B
Closed Hi-Hat	/t'/	t
PF Snare	/p'f:/	pf

3.1. Kick Drum

3.1.1. Closure

The kick drum was observed as a bilabial ejective affricate for all five subjects. Figure 1 contains images of the main postures for each subject during closure, release and after release of the main articulators during elicitation. Each series of images was taken over a single elicitation of the sound.

The mechanism for the production of the sound was similar for the two expert subjects. To prepare to make the sound, AM moves the tongue body back, lowers the larynx, raises the velum, and brings the lips together with a slight outward protrusion (Fig. 1A). This same preparation posture is seen for AF (Fig. 1D).

All five subjects exhibit raising of the velum and bilabial closure, while tongue body movement and lip protrusion vary. For IF and NM, the tongue body lowers instead of moving back (Fig. 1G, 1J). Additionally, for NM, the lips are protruding inward instead of outward (Fig. 1J). For NF, the tongue body does not lower as much as the experts', and the larynx does not lower either (Fig. 1M).

3.1.2. Release

The sound is made when air is forced through the lips. The velum stays raised for production of the sound, and the glottis is closed (Fig. 1B). The lip compression proceeds from closed to slightly apart, staying close enough to produce slight affrication. The sound is produced centrally for all five subjects; the midsagittal plane effectively captures the lip compression dynamics. High air pressure causes slight affrication after the initial stop, which adds a slight "punchy" quality to the sound.

For all of the subjects, with the exception of NF, a rapid upward movement of the larynx accompanies the sound production and lip release, indicating that this sound is an ejective. This larynx movement is evident by observing the difference in larynx heights between the "Closure" and "Release" columns of Figure 1. Closure of the glottis indicates that these four subjects use the glottalic egressive airstream for the production of this sound.

For NF, no closure or rapid raising of the larynx is observed (Fig. 1N). Therefore, no ejective is observed. As a result, the kick drum for NF sounds more similar to a "P" (SBN) as opposed to a "B" [2]. The "P" is a lighter and less "punchy" sound than the "B" [2].

3.1.3. Post-Release

The post-release postures for the kick drum vary across the subjects, indicating that the post-release posture is not crucial to the production of the sound.

After sound production, for both AM and AF the tongue body moves forward and the larynx moves back down to relax (Fig. 1C, 1F). For IF and NF, the mouth opens more. (Fig. 1I, 1O). For NM, the tongue body moves up (Fig. 1L).

Though not crucial to production of the sound, tongue position during post-release affects the pitch of the kick drum by changing the size of the vocal instrument [12].

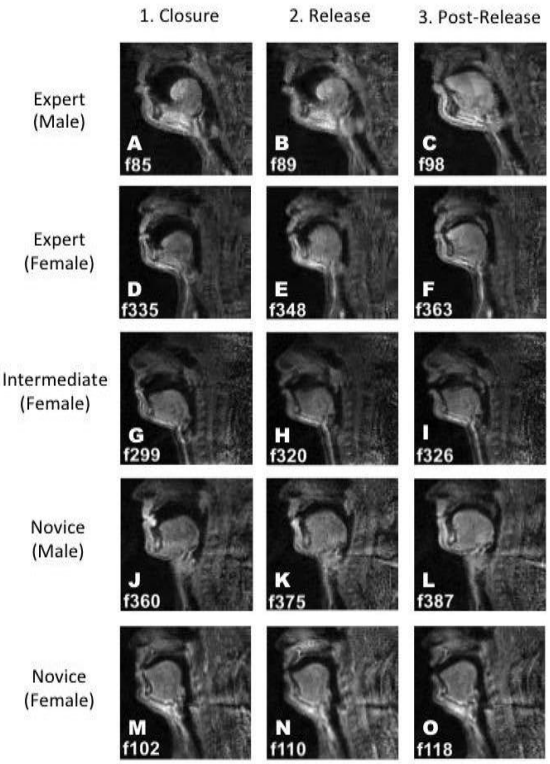

Figure 1: *Main postures found for elicitation of the "Kick Drum." Frame number from each MRI scan shown on bottom left corner of each image.*

3.2. Closed Hi-Hat

3.2.1. Closure

All five subjects produced the closed hi-hat as an alveolar affricate. Figure 2 shows the elicitation of this sound by AM and NM. To prepare to make the sound, the larynx lowers, the velum rises, the tongue body moves forward, and the tongue tip moves to the alveolar ridge (Fig. 2A)

3.2.2. Release

To produce the sound, the tongue tip releases from the alveolar ridge, with slight frication. NM uses the tongue front as opposed to the apex to make the alveolar closure (Fig. 2D). Yet, this produces no qualitative acoustical difference, suggesting that alveolar closure with either the tongue front or apex will accomplish the acoustical goal of this sound.

Simultaneously with this release, the larynx moves up to produce an ejective, as can be seen by the difference in larynx heights from the "Closure" to the "Release" postures (Fig. 2A, 2B). For NF, the ejective is less strong, but still present.

3.2.3. Post-Release

The tongue body moves to a more posterior position after the sound is made (Fig. 2C, 2F).

Acoustically, the sound is impressionistically similar across all subjects when heard outside the MRI machine. AM and NF have a more "breathy" sound, perhaps due to the use of a more pulmonic-heavy airstream after the initial alveolar closure and release.

The production of this sound may have been similar across all subjects since it is the closest of the three sounds to an articulation combination commonly found in spoken language [4].

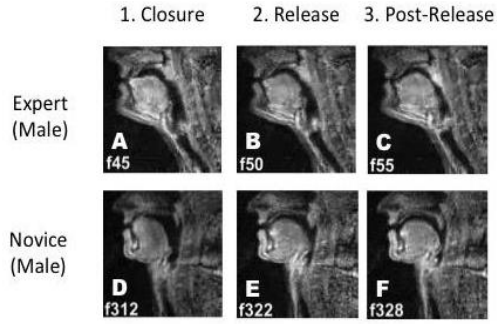

Figure 2: *Main postures for elicitation of the "Closed Hi-Hat"*

3.3. PF Snare

3.3.1. Closure

Of the three sounds studied in this paper, the mechanism of producing the PF snare varied the most across all five subjects. These differences are visible in Figure 3.

For AM, the pharynx widens, the velum raises, the tongue body becomes more compact and the lips come together with a slight outward protrusion to prepare for the sound (Fig. 3A).

AF prepares in a similar way, but with more puffing of the cheeks, suggesting more buildup of air pressure for creation of the sound (Fig. 3D). For IF, the tongue tip does not retract (Fig. 3G). For NM and NF, the tongue tip is positioned behind the teeth (Fig. 3J, 3M).

3.3.2. Release

This sound was produced either as a bilabial affricate ejective or as a bilabial stop. In both cases, a prolonged labiodental or bilabial fricative followed the initial articulation. The sound is produced when air is forced through the tightly compressed lips for the "P". The lips are more tightly compressed for the PF snare than they are for the kick drum across all five subjects.

For AF, IF and NM, the lips are released simultaneously with a rapid upwards motion of the larynx, showing an ejective (Fig. 3D, 3E). NF also exhibits raising of the larynx, though not as strongly (Fig. 3N), resulting in a less percussive sound.

For AM, there is no closure or raising of the larynx (Fig. 3B). Instead, he appears to use extra pulmonic pressure to achieve the sound quality characteristic of the PF snare. By doing so, AM shows that an ejective is not absolutely necessary for successful production of this sound.

For IF, the tongue body does not move as much as the experts (Fig. 3H). Additionally, the lips are not protruded as they were with the advanced beatboxers, resulting in a less "clean" sound. NF exhibits the same lack of outward lip

protrusion (Fig. 3N). As a result, IF and NF produce PF snares that are qualitatively less similar to the target sound, unlike the advanced beatboxers who produce a sound closely reminiscent of the corresponding snare on a drum set [4].

This failure to effectively mimic the physical instrument suggests that though IF and NF cannot produce the sound accurately, they are attempting to learn the sound. While learning to produce this sound, the novice and intermediate subjects attempt to accomplish articulations similar to the experts, but the exact placement of the lips and the timing is not coordinated the same way as the experts, indicating that the non-experts have less control over the articulations.

3.3.3. Post-Release

Similar to the kick drum, the post-release postures for the PF snare vary across the subjects, indicating that no particular post-release posture is crucial for production of the sound.

For the experts, the tongue body moves forward and follows through with the motion after the sound is produced (Fig. 3C, 3F). NM shows less tongue body movement than the experts (Fig. 3J-L).

For IF and NM, the lips do not come apart as much as they do for the experts (Fig. 3I, 3L). For NF, the lips come apart and the larynx relaxes (Fig. 3O).

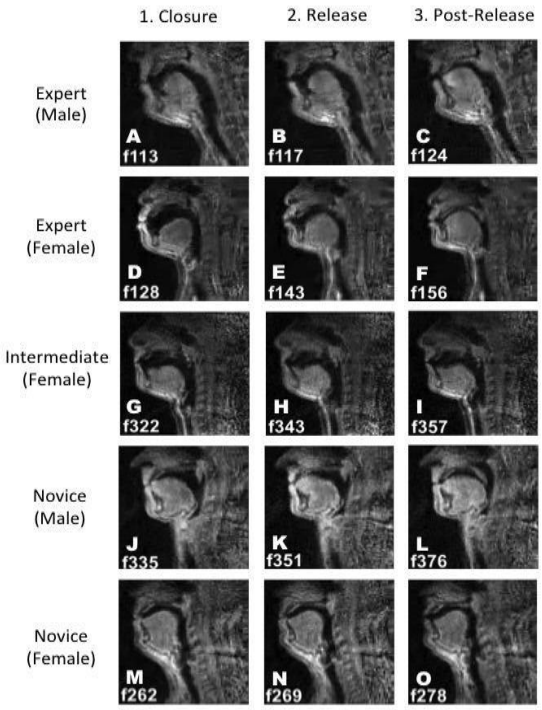

Figure 3: *Main postures for elicitation of the "PF Snare"*

4. Discussion

The articulatory data examined in this study offer some important insights into mechanisms of human sound production, airstream control, and ways in which the speech articulators may be recruited and coordinated for achieving musical, as well as linguistic goals.

By studying the production mechanisms across five speakers, the essential and non-essential postures for each sound can be determined. All five subjects raised the velum for these three sounds, suggesting that this articulatory technique is widely used in beatboxing. Furthermore, each of these three sounds had an oral closure that was necessary for the production of the sound across all five subjects: the bilabial closure for the kick drum, the alveolar closure for the closed hi-hat, and the tightly compressed bilabial closure for the PF snare.

Both males and females exhibited similar articulator movements when producing each sound. There was no discernible difference between genders, as the fundamentals of their vocal instruments were the same.

The variations in the post-release postures suggest that the posture following elicitation is not essential to the sound. This variation would allow for more flexibility in preparing for the next sound after one sound is elicited, which would help when using a sequence of sounds in a beat pattern.

Additionally, there was less variation amongst the sound closest to language (the closed hi-hat), while there was increasing variation with the sound that differed most from spoken language (the PF snare). This suggests that for the sounds further from those found in spoken language, our subjects appear to find different ways to manipulate their vocal instrument in order to produce the same sound.

Ejectives have more of a percussive quality, and stronger ejectives corresponded to a more percussive sound. In general, the novice female (NF) exhibited less strong ejectives, suggesting that novice beatboxers lack the laryngeal control exhibited by more advanced beatboxers. However, as exhibited by the advanced male's (AM) production of the PF snare, ejectives are not necessary for effective production of all sounds. In the particular case of AM's PF snare, stronger use of the pulmonic airstream compensated for the lack of the ejective, suggesting that different airstreams can be manipulated in different ways in order to affect and enhance the sound being produced.

While NF's lack of an ejective for the PF snare points to her lack of control over sound production, AM's lack of an ejective points to his artistic style. As beatboxing is primarily an art form, there is some stylistic variation between artists. NF fails to effectively mimic a drum set by not producing an ejective. AM compensates for the lack of an ejective by utilizing a pulmonic airstream leading to a determinedly "breathier" sound. This difference supports the theory that the goal of beatboxing is more acoustic as opposed to articulatory. This difference also supports the idea that differences in beatboxing skill levels require acoustic analysis in addition to quantitative articulatory analysis in order to examine and classify different styles in the art form.

5. Conclusions

Using rtMRI to compare beatboxing articulations between several subjects can shed light on the skill level and stylistic similarities and differences among beatboxers. While all artists exhibited the closures and releases essential for each sound, extraneous articulatory movements enhanced the quality and style of the sounds. These embellished articulations point to the possible production variations that provide the scientific basis behind the artistry of beatboxing.

6. Acknowledgements

This study was supported by NIH grant R01DC007124 and NSF grant IIS 1514544.

7. References

[1] TyTe and Defenicial, "Part 1: The Pre-History of Beatboxing", *Human Beatbox*, 2005. [Online]. Available: https://www.humanbeatbox.com/articles/history-of-beatboxing part-1/. [Accessed: 13- Mar- 2017].

[2] TyTe and Splinter, "Standard Beatbox Notation (SBN)", Human Beatbox, 2002. [Online]. Available: https://www.humanbeatbox.com/articles/standard-beatbox notation-sbn/. [Accessed: 20- Mar- 2017].

[3] D. Stowell and M. Plumbley, "Characteristics of the Beatboxing Vocal Style", *Human Beatbox*, 2008. [Online]. Available: https://www.humanbeatbox.com/articles/ characteristics-of-the-beatboxing-vocal-style/. [Accessed: 13- Mar- 2017].

[4] K. Lederer, "The Phonetics of Beatboxing", *Human Beatbox*, 2008. [Online]. Available: https://www.humanbeatbox.com/ articles/the-phonetics-of beatboxing-abstract/. [Accessed: 13- Mar- 2017].

[5] T. de Torcy, A. Clouet, C. Pillot-Loiseau, J. Vaissière, D. Brasnu and L. Crevier-Buchman, "A Video-Fiberscopic Study of Laryngopharyngeal Behaviour in the Human Beatbox", *Logopedics Phoniatrics Vocology*, vol. 39, no. 1, pp. 38-48, 2013.

[6] M. Proctor, E. Bresch, D. Byrd, K. Nayak and S. Narayanan, "Paralinguistic Mechanisms of Production in Human "Beatboxing": A Real-Time Magnetic Resonance Imaging Study", *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 1043-1054, 2013.

[7] Stone, M. et al. "Comparison of Speech Production in Upright and Supine Position". The Journal of the Acoustical Society of America 122.1 (2007): 532-541. Web. 5 June 2017.

[8] S. Lingala, Y. Zhu, Y. Kim, A. Toutios, S. Narayanan and K. Nayak, "A fast and flexible MRI system for the study of dynamic vocal tract shaping", Magnetic Resonance in Medicine, vol. 77, no. 1, pp. 112-125, 2016.

[9] S. Narayanan, K. Nayak, S. Lee, A. Sethy and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production", The Journal of the Acoustical Society of America, vol. 115, no. 4, pp. 1771-1776, 2004.

[10] "Berkeley Advanced Reconstruction Toolbox", GitHub, 2017. [Online]. Available: https://mrirecon.github.io/bart/. [Accessed: 21- Mar- 2017].

[11] E. Bresch, J. Nielsen, K. Nayak and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans", The Journal of the Acoustical Society of America, vol. 120, no. 4, pp. 1791-1794, 2006.

[12] "Factors Influencing Fundamental Frequency", National Center for Voice and Speech, 2017. [Online]. Available: http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/ influence.html. [Accessed: 21- Mar- 2017].