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

Sounds of the Human Vocal Tract

Reed Blaylock, Nimisha Patil, Timothy Greer, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory
University of Southern California, USA

reed.blaylock@gmail.com, {nimishhp, timothdg}@usc.edu, shri@sipi.usc.edu

Abstract

Previous research suggests that beatboxers only use sounds that exist in the world's languages. This paper provides evidence to the contrary, showing that beatboxers use non-linguistic articulations and airstream mechanisms to produce many sound effects that have not been attested in any language. An analysis of real-time magnetic resonance videos of beatboxing reveals that beatboxers produce non-linguistic articulations such as ingressive retroflex trills and ingressive lateral bilabial trills. In addition, beatboxers can use both lingual egressive and pulmonic ingressive airstreams, neither of which have been reported in any language.

The results of this study affect our understanding of the limits of the human vocal tract, and address questions about the mental units that encode music and phonological grammar.

Index Terms: beatboxing, MRI, singing, speech production

1. Introduction

Although the human vocal tract is most well-known and widely studied for its speech movements, it also produces skilled movements in domains other than language. Many of those skilled movements are unattested in languages used today. For instance, throat singers and metal singers use epilaryngeal vibration to create a growl-like vocalization [1]. This growl appears in speech as a marker of emotion, but is not contrastive in any known phonologies. Similarly, some operatic singers produce a singer's formant (a clustering of the third, fourth, and fifth formants) to ensure that the sound of their voice can be heard over an orchestra [2]. The production of the singer's formant is the result of precise articulatory positioning that has no known role in speech. This paper describes the articulatory patterns used in beatboxing, a type of musical expression that requires skilled movements of the vocal tract—many of which are not found in known languages.

In beatboxing, the human vocal tract is used to emulate percussion and other physical or synthesized sounds. Vocal percussion is a component of many musical styles around the world, including Southern Indian Carnatic musical traditions and North American scat singing [3]. In the past few decades, beatboxing has been used in hip hop and a cappella music, and as a stand-alone musical style [4].

Beatboxing is of particular interest to speech scientists because it uses speech sounds or speech-like sounds to create music that often has no linguistic content. In beatboxing, percussion sounds like kick drums, hi-hats, and snares can be emulated by the vocal tract with bilabial ejectives, alveolar affricates, and trills—all of which appear in language.

Given the apparent similarities between beatboxing and speech, questions like the following emerge:

- Do speech and beatboxing articulations share the same mental representations?
- Is beatboxing grammatical in the same way that phonology is grammatical?
- How is the musical component of beatboxing represented in the mind?

To address these questions, a comprehensive inventory of beatboxing sounds and their articulations must first be compiled.

Some steps toward this goal have already been taken. Splinter and TyTe [5] proposed a Standard Beatbox Notation (SBN) to encode contrasts between different beatboxing sounds, and Stowell [6] uses a combination of new characters and characters from the International Phonetic Alphabet (IPA) to represent some sounds of beatboxing. Stowell and Plumbley [7] and Stowell [8] impressionistically describe the acoustic properties of certain beatboxing sounds compared to speech and singing sounds. Lederer [9] provides a detailed acoustical analysis of a few common beatboxing sounds, and explains how to produce those sounds based on descriptions provided by the study's subject. All of these studies, however, are based on impressionistic and acoustic data; none examine the articulatory strategies used in beatboxing.

In an articulatory study using real-time magnetic resonance imaging (rtMRI), Proctor et al. [10] reported the productions of one beatboxer whose sounds were all produced with pulmonic egressive, glottalic egressive ("ejective"), or lingual ingressive ("click") airstreams, and had articulations amenable to transcription using IPA—an orthographic system that was designed to describe speech sounds that are contrastive in languages of the world.

As Proctor et al. point out, their study is limited by having only one beatboxer; this beatboxer apparently did not have any non-linguistic sounds in their repertoire. Further observation of skilled beatboxers, including the acoustic research cited above, clearly indicates that beatboxers deploy articulator strategies that are not represented in the sounds of the world's languages. What those strategies are, however, is currently unknown.

The aim of this paper is to show that beatboxers use non-linguistic articulations in addition to linguistic (or pseudo-linguistic) productions. In particular, this paper highlights the articulatory configurations and airstream mechanisms that are unattested in language, which until now have not been described for beatboxing. The findings of this paper point to further avenues of investigation for understanding skilled vocal tract movement in different domains, including the possibility of a phonotactic grammar for beatboxing.

MR videos of the sounds in this paper can be found at http://sail.usc.edu/span/beatboxing2017/index.html.

http://dx.doi.org/10.21437/Interspeech.2017-1631

2. Method

Five beatboxers were asked to produce beatboxing sounds in isolation and in musical rhythms (“beats”), and to speak several passages, in an MRI machine. Of those five participants, the productions of just two—participants IF (second author of this paper) and AF, both females—are reported in the present study. Participant IF was an intermediate-level beatboxer, and participant AF was an advanced-level beatboxer. Skill level was defined based on the participant’s level of artistic control and the apparent difficulty of the sounds in their repertoire. Both participants reported English as their native language, and IF also reported fluency in Marathi and Hindi. Of the sounds reported, only some were produced by both IF and AF.

The participants were asked to produce all of the percussion effects in their repertoire and demonstrate some beatboxing sequences by performing in short intervals as they lay supine in an MRI scanner bore. For each sound the elicitation was repeated at least three times in a single MRI recording and subsequently used in a sample beat pattern. In addition, some speech passages were recorded, and a full set of the subject's American English vowels was elicited using the [h_d] corpus. The subjects were paid for participation in the experiment. The study presented in this paper draws from a subset of this data.

Data were collected using an rtMRI protocol developed for the dynamic study of vocal tract movements, especially during speech production [11, 12]. The subjects’ upper airways were imaged in the midsagittal plane using a gradient echo pulse sequence (TR = 6.004 ms) on a conventional GE Signa 1.5 T scanner (Gmax = 40 mT/m; Smax = 150 mT/m/ms), using an 8-channel upper-airway custom coil.

The slice thickness for the scan was 6 mm, located midsagittally over a 200 mm × 200 mm field-of-view; image resolution in the sagittal plane was 84 × 84 pixels (2.4 × 2.4 mm). The scan plane was manually aligned with the midsagittal plane of the subject’s head. The frames were retrospectively reconstructed to a temporal resolution of 12ms (2 spirals per frame, 83 frames per second) using a temporal finite difference constrained reconstruction algorithm [12] and a recent open-source library [13].

Audio was recorded at a sampling frequency of 20 kHz inside the MRI scanner while the subjects were imaged, using a custom fiber-optic microphone system. The audio recordings were noise-canceled, then reintegrated with the reconstructed MR-imaged video [14]. The result allows for dynamic visualization and synchronous audio of the performers’ vocal tracts. Because the scan plane was in the midsagittal plane of the glottis, it was possible to observe glottal abduction and adduction, as well as vertical laryngeal movements.

3. Sounds

3.1. Linguistic sounds

Proctor et al. [10] showed that a selection of basic beatboxing sounds correspond to a set of linguistic sounds. Many of the utterances in the present study can also be described using IPA, and match sounds attested in various languages of the world.

Table 1 describes some of the linguistic sounds produced by IF and AF. The representation of each sound in IPA is provided, as well as a full featural description.

The kick drum, closed hi-hat, and PF snare are some of the standard beatboxing sounds; as they stem from the old-school style of beatboxing and are comparable to sounds made in language, they are among the first for new beatboxers to learn, and they are used frequently at all levels of beatboxing [4, 5].

Table 1: *Sample of beatboxing sounds that appear linguistic*

Name	IPA	Description
Kick drum, “B”	[p͡ɸ’]	Voiceless glottalic egressive (ejective) bilabial affricate
Closed hi-hat	[ts’]	Voiceless glottalic egressive (ejective) alveolar affricate
PF snare, “PF”	[pf’:]	Voiceless glottalic egressive (ejective) labial affricate (with long frication)
Voiced tongue bass	[r̻]	Voiced pulmonic egressive laminal alveolar trill

3.2. Non-linguistic sounds

3.2.1. Non-linguistic airflow

Though some of the sounds reported here are similar to sounds used in language, other beatboxing sounds were also observed that are not attested in any language known to the authors.

One source of novelty is that beatboxers use a wider range of airstream mechanisms. Speech relies almost exclusively on four airstreams: pulmonic egressive (e.g. [k], glottalic egressive (e.g. [p’]), glottalic ingressive (e.g. [ɓ]), and lingual ingressive (e.g. [ʘ]). Previous studies in beatboxing have differentiated sounds as using “inhalation” or “exhalation,” but have not specified whether the airflow was due to glottal, lingual, or pulmonic activity [5, 7, 8, 9]. It has been reported that beatboxers use all airstreams found in language, but does not report on whether they use non-linguistic airstreams [9]. This study finds that beatboxers can—and frequently do—use non-linguistic pulmonic ingressive and lingual egressive airstreams.

Pulmonic ingressive airflow is inhalation. Inhalation is not used for linguistic content in speech, but it is a crucial preparatory step for pulmonic egression (exhalation). Audible inhalation (i.e. gasping), accomplished by ingressive breathy voicing, communicates acute onset of emotion, such as shock or excitement. Pulmonic ingressive airflow is sometimes used for certain interjections or phrases, but does not appear to be phonologically contrastive in any language [15].

Lingual egressive airflow is the opposite of a click (which would be lingual ingressive). In a click, the tongue body makes a constriction against the palate or velum, and the lips or tongue tip make another constriction in front of the tongue body constriction, creating a seal. The tongue then moves backward, increasing the volume and decreasing the pressure within the seal. When the anterior constriction is released, air is pulled toward the space that was formerly the low pressure seal, resulting in a click’s characteristic smacking or popping sound.

For a lingual egressive sound, two constrictions are made, just as in a click: one constriction of the tongue body against the velum or palate, and one constriction of the tongue tip or lips. But rather than pulling the tongue body backward, as in a click, the tongue body moves forward toward the anterior constriction, increasing air pressure and forcing the air outward. Figure 1 shows two stages of the Lip Pop, a lingual egressive sound produced by first making closures with the lips and tongue body (f254), then by using the tongue to push air through the lips with a quick "pop" (f267). Labels (e.g. f254) indicate the frame number for the video of the observed sound.

Lingual egressive sounds are not found in language [15], but a similar mechanism is used for circular breathing, a technique used by some wind instrument players to sustain sound for longer than one breath would allow. In circle breathing, a seal is made by the tongue body against the velum or palate while the cheeks are expanded; by compressing the cheeks, musicians can squeeze air into their instrument while simultaneously inhaling to replenish their breath.

Beatboxers can use pulmonic ingressive and lingual egressive airstreams with a wide variety of constriction locations and degrees of constriction. The lingual egressive airstream can be used to create (at least) trills and stop-like sounds. The pulmonic ingressive airstream is used for fricatives and trills, and possibly more. A sample list of the beatboxing sounds that use non-linguistic airstreams is given in Table 2.

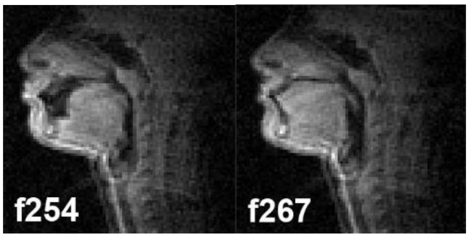

Figure 1: *IF producing a lip pop. The lips and tongue body make closures (left), then the tongue pushes forward (right)*

Table 2: *Sample of beatboxing sounds that use non-linguistic airflow*

Name	Description
(Inward) K	Voiceless pulmonic ingressive lateral velar fricative
Snore bass	Voiceless pulmonic ingressive uvular trill
Zipper	Voiceless pulmonic ingressive bilabial fricative
Lip pop	Voiceless lingual egressive labial stop
BMG/Spit snare	Voiceless buccal-lingual egressive labial affricate
Forced hi-hat	Voiceless lingual egressive alveolar affricate
Clickroll	Voiceless lingual egressive alveolar trill

3.2.2. Non-linguistic articulations

Beatboxers can also produce sounds whose vocal tract configurations are non-linguistic (Table 3).

The liproll, for example, is a voiceless lingual ingressive lateral bilabial trill. In language, labial trills are made with pulmonic egressive airstream rather than lingual ingressive airstream. Equally ununusual is the laterality of this trill, which uses either the left or right side of the lips in isolation. In language, laterality is exclusively used to describe tongue shape; a lateral labial articulation is not found in speech. The liproll can also be produced with pulmonic ingressive airflow, and with simultaneous tongue movement to affect sound quality.

One remarkable variation of the liproll is the inward clickroll with liproll (Figure 2). The inward clickroll is a pulmonic ingressive retroflex trill. When combined with a liproll, the result is two simultaneous pulmonic ingressive trills—one labial, and one retroflex.

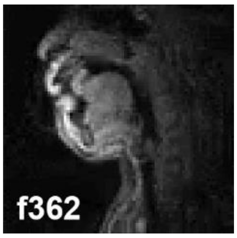

Figure 2: *AF producing an inward clickroll with liproll*

Table 3: *Sample of beatboxing sounds that use non-linguistic articulations*

Name	Description
Liproll	Voiceless lingual egressive lateral bilabial trill
Inward clickroll	Voiceless pulmonic ingressive retroflex trill

4. Discussion

Given the breadth of beatboxing sounds observed in this study, it is clear that beatboxers use a variety of sounds, some of which are attested in language, and many of which are not attested in any known language.

Section 4.1 discusses the extent to which linguistic features can be used to describe beatboxing sounds. Section 4.2 identifies linguistic gaps in beatboxing, and section 4.3 considers the implications of these gaps in the context of a phonotactic grammar of beatboxing. Section 4.4 addresses possible articulatory differences between linguistic beatboxing sounds and the corresponding sounds used in language.

4.1. Limits of describing beatboxing using linguistic features

Overall, the beatboxing sounds reported here were described using pre-existing terminology drawn from a common set of linguistic features. The sounds made with non-linguistic airstreams used "pulmonic ingressive" and "lingual egressive" features; while those terms are not linguistic (because those airstreams do not appear in language), their components ("pulmonic," "lingual," "ingressive," "egressive") are commonly used. Likewise, non-linguistic articulations such the liproll's "lateral bilabial trill" are made of the well-attested features "lateral," "bilabial," and "trill."

One exception is the BMG/spit snare, which the intermediate-level beatboxer IF produced with an unusual airstream and manner of articulation. Air was forced out of the mouth using compression of the cheeks; the tongue was only minimally involved. To our knowledge, cheek compression is not a contrastive linguistic feature, and not a component of IPA. The reported airstream is buccal-lingual egressive, though emphasis should be on the buccal (cheek) component.

The reported manner of articulation for the BMG/spit snare is "affricate," which implies a combination of a stop and fricative. While the lips do compress against each other to stop the air, there is no subsequent release during this sound as with a typical stop or affricate; instead, the air is squeezed through the closed lips. To the authors' knowledge, there is no feature that matches this type of airflow.

Most of the sounds reported here accommodated a description using linguistic features, though some features were used unconventionally. There are, however, more possible beatboxing sounds than the ones reported here, some of which may not fit into the same terminology.

4.2. Language gaps in the beatboxing inventory

The beatboxing sounds reported here cover six airstream mechanisms, four linguistic and two non-linguistic, as well as a variety of articulatory postures and movements. However, there are many linguistic sounds whose articulations were not produced in this study. For instance, IF produced a voiceless pulmonic ingressive uvular trill ("snore bass"), but did not produce the closest linguistic correspondent to that sound—a voiced pulmonic egressive uvular trill. Also absent are certain nasals, pharyngeals, approximants, palatals, many fricatives, most voiced sounds, plain stops, and vowels.

Presumably beatboxers—who spend hours developing and learning articulatory combinations that are wildly un-speechlike—are also capable of using sounds from their native language to emulate percussion or other instruments. Instead, it appears that beatboxers deliberately remove some speech sounds from their beatboxing inventory, rather than trying to use sounds from their native language.

Plain stops (e.g. [p], [k]) are often the first to go. A typical first attempt at beatboxing is to repeat the phrase "boots and cats and…" As beatboxers become more skilled, the vowels in these words disappear, and the consonants become more aggressive. The [b] of "boots" becomes an ejective kick drum, the [ts] of "boots" and "cats" becomes a hi-hat, and the [k] of "cats" becomes an inward or outward K. Our companion paper provides evidence for this [16]: while advanced and intermediate beatboxers consistently produce ejectives for their kick drums and closed hi-hats, novice beatboxers are more likely to produce those sounds as plain stops.

It appears that skilled beatboxers learn to avoid plain stops in favor of typologically rarer ejectives. Similar processes of speech sound removal likely apply to other sounds like vowels.

4.3. Beatboxing phonotactics

Given the evidence from section 4.2, it appears that there are restrictions in beatboxing that determine which sounds are permissible.

In the speech domain, this set of restrictions would be called a language's phonotactics, and would be part of a language's phonological grammar. For example, most languages strongly prefer voiced sonorants (e.g. [m], [r]) to voiceless sonorants (e.g. [m̥], [r̥]. Voiceless sonorants are possible to produce, but they are nearly inaudible. Since a primary function of language is communication, and communication by sound requires that sounds are loud enough to be heard, one could say that "audibility" is a general requirement for speech sounds. This audibility restriction imposes a phonotactic restriction on voiceless sonorants.

Extending this logic, one could say that beatboxing has a similar general goal of "being percussive." Sounds that are less percussive, like plain stops, are dispreferred in beatboxing, just like voiceless sonorants are dispreferred in speech.

As an alternative to proposing beatboxing phonotactics, one could argue that the apparent restriction against some linguistic sounds is the result of musicality: the more high-pressure sounds are traditionally popular, and just happen to sound more like the instruments that are the subjects of mimicry.

These two approaches may in fact be compatible with each other; but, at first glance, they make distinct predictions. The phonotactic hypothesis predicts that beatboxing will exhibit other sound patterns, similar to patterns that appear in language. For instance, some languages exhibit long-distance harmony, where one or more qualities of a sound propagate to other compatible sounds later in the word. Beatboxing may exhibit a similar harmony in musical phrases. Some languages also restrict certain sounds from occurring with other sounds, such as two adjacent vowels ("hiatus") or several consonants of relatively equal sonority. Beatboxing phrases could exhibit these types of restrictions as well.

An extension of the phonotactic hypothesis would suggest that the different domains of skilled vocal tract movement all have grammatical components, similar to the grammar of speech.

The musicality hypothesis predicts no such restrictions. If no phonotactic grammar exists for beatboxing, then beatboxers should be able to combine any sound with any others, with no limitations other than personal taste. These predictions are empirically testable with a corpus of beatboxing phrases, such as the collection of rtMR videos collected as part of this study.

4.4. What it means to be "linguistic"

In describing some beatboxing sounds as "linguistic," there is an implicit assumption that the production of those sounds matches the productions used in language. For example, the characterization of the kick drum as an ejective is based on its quick laryngeal raising near the bilabial release. However, it is unknown if this larynx raising matches timing and magnitude of larynx raising observed in language. Beatboxing seems to require more highly-pressurized productions than speech, especially in ejectives. Therefore, one might expect that beatboxing ejectives have larger larynx movements than speech ejectives, or different timing between lip and glottis closure. Even though some beatboxing sounds can be described using attested combinations of features, the actual articulatory strategies used may differ from any strategies in language.

It is also important to remember that sounds that are not used in any contemporary language may have been used in language in the past, or may be used in language in the future. For now, it is sufficient to say that beatboxers can use sounds that have no linguistic counterpart in their own languages.

5. Conclusions

This paper has qualitatively shown, using rtMRI video data, that beatboxers use a combination of linguistic and non-linguistic sounds in their musical production. In most cases, the sounds reported fit a description using linguistic features.

For fundamental beatboxing sounds, beatboxers favor ejectives over plain stops; in general, the standard inventory of beatboxing sounds is substantially different from the set of sounds that are common in language (e.g. stops, vowels). Beatboxing may have its own phonotactic grammar that restricts which sounds are permissible in vocal percussion.

6. Acknowledgements

This work is supported by NIH grant R01DC007124 and NSF grant IIS 1514544.

7. References

[1] S. R. Moisik, “The epilarynx in speech,” Ph.D. dissertation, Dept. Ling., Univ. Victoria, 2013.

[2] J. Sundberg, “Articulatory interpretation of the “singing formant”.” *J. Acoust. Soc. Amer*., vol. 55, no. 4, pp. 838-844, 1974.

[3] M. Atherton, “ Rhythm-speak: Mnemonic, language play or song,” in *Proc. Inaugural Int. Conf. Music Communication Science (ICoMCS)*, Sydney, edited by E. Schubert *et al.*., pp. 15–18, 2007.

[4] G. TyTe and Defenicial. (2005). *Part 1: The Pre-History of Beatboxing* [Online]. Available: https://www.humanbeatbox.com/articles/history-of-beatboxing-part-1/.

[5] M. Splinter and G. TyTe. (2002). *Standard Beatbox Notation (SBN)* [Online]. Available: https://www.humanbeatbox.com/articles/standard-beatbox-notation-sbn/.

[6] D. Stowell. (2008–2012). *The beatbox alphabet*. Available: http://www.mcld.co.uk/beatboxalphabet/.

[7] D. Stowell and M. Plumbley. (2008). *Characteristics of the Beatboxing Vocal Style* [Online]. Available: https://www.humanbeatbox.com/articles/characteristics-of-the-beatboxing-vocal-style/.

[8] D. Stowell, “Making music through real-time voice timbre analysis: machine learning and timbral control,” Ph.D. dissertation, School Elec. Eng. and Comp. Sci., Queen Mary Univ., London, 2010.

[9] Lederer. (2006). *The phonetics of beatboxing* [Online]. Available: https://www.humanbeatbox.com/articles/the-phonetics-of-beatboxing-part-4/.

[10] M. Proctor, et al., "Paralinguistic mechanisms of production in human “beatboxing”: A real-time magnetic resonance imaging study," *J. Acoust. Soc. Amer*., vol. 133, no. 2, pp. 1043-1054, 2013.

[11] S. Narayanan et al., “An approach to realtime magnetic resonance imaging for speech production,” *J. Acoust. Soc. Amer*., vol. 115, no. 4, pp. 1771–1776, 2004.

[12] S. G. Lingala et al., “A Fast and Flexible MRI System for the Study of Dynamic Vocal Tract Shaping,” *Magentic resonance in medicine*, 2016.

[13] BART Reconstruction [Online]. Available: https://mrirecon.github.io/bart/.

[14] E. Bresch et al., “Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP],” *IEEE Signal Process. Mag*, vol. 25, no. 3, pp. 123–132, 2008.

[15] R. Eklund, “Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in humans speech,” J. Intl. Phonetic Assoc., vol. 38, no. 3, 2008.

[16] N. Patil et al., “Comparison of Basic Beatboxing Articulations Between Expert and Novice Artists using Real-Time Magnetic Resonance Imaging,” INTERSPEECH, 2017.