Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study

Michael Proctor[a)]
Viterbi School of Engineering, University of Southern California, 3740 McClintock Avenue, Los Angeles, California 90089-2564

Erik Bresch
Philips Research, High Tech Campus 5, 5656 AE, Eindhoven, Netherlands

Dani Byrd
Department of Linguistics, University of Southern California, 3601 Watt Way, Los Angeles, California 90089-1693

Krishna Nayak and Shrikanth Narayanan
Viterbi School of Engineering, University of Southern California, 3740 McClintock Avenue, Los Angeles, California 90089-2564

(Received 6 March 2012; revised 30 October 2012; accepted 17 December 2012)

Real-time Magnetic Resonance Imaging (rtMRI) was used to examine mechanisms of sound production by an American male beatbox artist. rtMRI was found to be a useful modality with which to study this form of sound production, providing a global dynamic view of the midsagittal vocal tract at frame rates sufficient to observe the movement and coordination of critical articulators. The subject's repertoire included percussion elements generated using a wide range of articulatory and airstream mechanisms. Many of the same mechanisms observed in human speech production were exploited for musical effect, including patterns of articulation that do not occur in the phonologies of the artist's native languages: ejectives and clicks. The data offer insights into the paralinguistic use of phonetic primitives and the ways in which they are coordinated in this style of musical performance. A unified formalism for describing both musical and phonetic dimensions of human vocal percussion performance is proposed. Audio and video data illustrating production and orchestration of beatboxing sound effects are provided in a companion annotated corpus.
 [http://dx.doi.org/10.1121/1.4773865]

PACS number(s): 43.70.Bk, 43.75.St, 43.70.Mn, 43.75.Rs [BHS] Pages: 1043–1054

I. INTRODUCTION

Beatboxing is an artistic form of human sound production in which the vocal organs are used to imitate percussion instruments. The use of vocal percussion in musical performance has a long history in many cultures, including *konnakol* recitation of *solkattu* in Karnatic musical traditions of southern India, North American *a capella* and *scat singing*, Celtic *lilting* and *diddling*, and Chinese *kouji* performances (Atherton, 2007). Vocal emulation of percussion sounds has also been used pedagogically, and as a means of communicating rhythmic motifs. In north Indian musical traditions *bols* are used to encode tabla rhythms; *changgo* drum notation is expressed using vocables in Korean *samul nori*, and Cuban conga players vocalize drum motifs as *guauganco* or *tumbao* patterns (Atherton, 2007; McLean and Wiggins, 2009).

In contemporary western popular music, human beatboxing is an element of hip hop culture, performed either as its own form of artistic expression, or as an accompaniment to rapping or singing. Beatboxing was pioneered in the 1980s by New York artists including Doug E. Fresh and Darren Robinson (Hess, 2007). The name reflects the origins of the practice, in which performers attempted to imitate the sounds of the synthetic drum machines that were popularly used in hip hop production at the time, such as the TR-808 Rhythm Composer (Roland Corporation, 1980) and the LM-1 Drum Computer (Linn Electronics, 1982). Artists such as Biz Markie, Rahzel, and Felix Zenger have advanced the art form by extending the repertoire of percussion sounds that are emulated, the complexity of the performance, and the ability to create impressions of polyphony through the integrated production of percussion with a bass line or sung lyrics.

Because it is a relatively young vocal art form, beatboxing has not been extensively studied in the musical performance or speech science literature. Acoustic properties of some of the sounds used in beatboxing have been described impressionistically and compared to speech sounds (Stowell and Plumbley, 2008). Stowell (2010, 2012) and Tyte (2012) have surveyed the range of sounds exploited by beatbox artists and the ways in which they are thought to be commonly produced. Splinter and Tyte (2012) have proposed an informal system of notation (Standard Beatbox Notation, SBN), and Stowell (2012) has outlined a modified subset of the

a)Author to whom correspondence should be addressed. Current address: MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith NSW 2751, Australia. Electronic mail: michael.proctor@uws.edu.au

International Phonetic Alphabet (IPA) to describe beatbox performance, based on these assumptions.

Lederer (2005) conducted spectral analyses of three common effects produced by human beatbox artists, and compared these, using 12 acoustic metrics, to equivalent electronically generated sounds. Sinyor *et al*. (2005) extracted 24 acoustic features from recordings of five imitated percussion effects, for the purpose of automatic categorization. Stowell and Plumbley (2010) examined real-time classification accuracy of an annotated dataset of 14 sounds produced by expert beatboxers. Acoustic feature analysis of vocal percussion imitation by non-beatboxers has also been conducted in music retrieval systems research (e.g., Kapur *et al*., 2004).

Although these studies have laid some foundations for formal analysis of beatboxing performance, the phonetics of human-simulated percussion effects have not been examined in detail. It is not known to what extent beatbox artists use the same mechanisms of production as those exploited in human language. Furthermore, it is not well understood how artists are able coordinate linguistic and paralinguistic articulations so as to create the perception of multiple percussion instruments, and the illusion of synchronous speech and accompanying percussion produced by a single performer.

II. GOALS

The goal of the current study is to begin to formally describe the articulatory phonetics involved in human beatboxing performance. Specifically, we make use of dynamic imaging technology to

(1) document the range of percussion sound effects in the repertoire of a beatbox artist;
(2) examine the articulatory means of production of each of these elements;
(3) compare the production of beatboxing effects with similar sounds used in human languages; and
(4) develop a system of notation capable of describing in detail the relationship between the musical and phonetic properties of beatboxing performance.

Through detailed examination of this highly specialized form of vocal performance, we hope to shed light on broader issues of human sound production—making use of direct articulatory evidence to seek a more complete description of phonetic and artistic strategies for vocalization.

III. CORPORA AND DATA ACQUISITION

A. Participant

The study participant was a 27 year-old male professional singer based in Los Angeles, CA. The subject is a practitioner of a wide variety of vocal performance styles including hip hop, soul, pop, and folk. At the time of the study, he had been working professionally for 10 years as an *emcee* (rapper) in a hip hop duo, and as a session vocalist with other hip hop and fusion groups. The subject was born in Orange County, CA, to Panamanian parents, is a native speaker of American English, and a heritage speaker of Panamanian Spanish.

B. Corpus

The participant was asked to produce all of the percussion effects in his repertoire and to demonstrate some beatboxing sequences, by performing in short intervals as he lay supine in an MRI scanner bore. Forty recordings were made, each lasting between 20 and 40 s, of a variety of individual percussion sounds, composite beats, rapped lyrics, sung lyrics, and freestyle combinations of these elements. In addition, some spontaneous speech was recorded, and a full set of the subject's American English vowels was elicited using the [h_d] corpus. The subject was paid for his participation in the experiment.

Individual percussion sounds were categorized by the subject into five instrumental classes: (1) kick drums, (2) rim shots, (3) snare drums, (4) hi-hats, and (5) cymbals (Table I, column 1). Further descriptions were provided by the subject in English to describe the specific percussion effect being emulated (Table I, column 2). For each demonstration the target effect was repeated at least five times in a single MRI recording, with elicitations separated by short pauses of approximately 2 s.

Each repeatable rhythmic sequence, or "groove," was elicited multiple times at different tempi, ranging from slow [approximately 88 beats per minute (b.p.m.)] to fast (~104 b.p.m.). The subject announced the target tempo before producing each groove and paced himself without the assistance of a metronome or any other external stimuli.

C. Image and audio acquisition

Data were acquired using a real-time Magnetic Resonance Imaging (rtMRI) protocol developed specifically for the dynamic study of upper airway movements, especially during speech production (Narayanan *et al*., 2004). The subject's upper airway was imaged in the midsagittal plane using a gradient echo pulse sequence ($T_R = 6.856$ ms) on a

TABLE I. Musical classification and phonetic characterizatioan of beatboxing effects in the repertoire of the study subject.

Effect	Description	SBN	IPA	Airstream
Kick	"punchy"	bf	[p͡f'ːʌ̹]	glottalic egressive
Kick	"thud"	b	[p'ɪ̹]	glottalic egressive
Kick	"808"	b	[p'ʊ̹]	glottalic egressive
Rimshot		k	[k']	glottalic egressive
Rimshot	"K"	k	[k^{h}hː]	pulmonic egressive
Rimshot	"side K"		[ɴ̹‖]	lingual ingressive
Rimshot	"sucking in"		[ɴ̹!]	lingual ingressive
Snare	"clap"		[ɴ̹|^{w}]	lingual ingressive
Snare	"no meshed"	pf	[p͡f'ːʉ̹]	glottalic egressive
Snare	"meshed"	ksh	[kç̻ː]	pulmonic egressive
Hi-hat	"open K"	kss	[ksː]	pulmonic egressive
Hi-hat	"open T"	tss	[t̻͡s̻ː]	pulmonic egressive
Hi-hat	"closed T"	^t	[t̻͡s̻̚]	pulmonic egressive
Hi-hat	"kiss teeth"	th	[^{w}ɴ̹‖]	lingual ingressive
Hi-hat	"breathy"	h	[xːʷ]	pulmonic egressive
Cymbal	"with a T"	tsh	[tɕːʷ]	pulmonic egressive
Cymbal	"with a K"	ksh	[k^{w}ɕːʷ]	pulmonic egressive

conventional GE Signa 1.5 T scanner ($G_{max} = 40$ mT/m; $S_{max} = 150$ mT/m/ms), using a generic 4-channel head-and-neck receiver coil.

Scan slice thickness was 5 mm, located midsagittally over a 200 mm × 200 mm field-of-view; image resolution in the sagittal plane was 68 × 68 pixels (2.9 × 2.9 mm). MR image data were acquired at a rate of 9 frames per second (f.p.s.), and reconstructed into video sequences with a frame rate of 20.8 f.p.s. using a gridding reconstruction method (Bresch *et al.*, 2008).

Audio was simultaneously recorded at a sampling frequency of 20 kHz inside the MRI scanner while the subject was imaged, using a custom fiber-optic microphone system. Audio recordings were subsequently noise-canceled, then reintegrated with the reconstructed MR-imaged video (Bresch *et al.*, 2006). The resulting data allows for dynamic visualization, with synchronous audio, of the performer's entire midsagittal vocal tract, from the upper trachea to the lips, including the oropharynx, velum, and nasal cavity. Because the scan plane was located in the midsagittal plane of the glottis, abduction and adduction of the vocal folds could also be observed.

IV. DATA ANALYSIS

Companion audio and video recordings were synchronized and loaded into a custom graphic user interface for inspection and analysis (Proctor *et al.*, 2010a; Narayanan *et al.*, 2011), so that MR image sequences could be examined to determine the mechanisms of production of each of the sound effects in the subject's repertoire.

Start and end times delineating each token were identified by examining the audio signal, spectrogram, and time-aligned video frames, and the corresponding intervals of each signal were labeled. Laryngeal displacement was calculated by manually locating the end points of the glottal trajectory using a measurement cursor superimposed on the video frames. The coordination of glottal and supraglottal gestures was examined to provide insights into the airstream mechanisms exploited by the artist to produce different effects.

Beatboxing grooves produced by the subject were manually transcribed. Using MuseScore (v1.2) musical notation software, the proposed transcriptions were encoded in MIDI format, exported as WAV audio, and compared to the audio recordings of the corresponding performance segment. To ensure that the annotated percussion sequences captured the musical properties of the grooves performed by the subject as accurately as possible, the musical scores and specifications for percussion ensemble, tempo and dynamics were adjusted, along with the MIDI sound palates, until the synthesized audio closely approximated the original recordings.

V. RESULTS

Seventeen phonetically distinct percussion effects occurred in this performer's repertoire, summarized in Table I.[1] For each sound, the performer's own description of the percussion class and intended effect is listed first, followed by a description in Standard Beatbox Notation, where this exists, using the conventions proposed by Splinter and Tyte (2012). IPA transcriptions of the articulatory configuration observed during each effect are proposed in column 4, along with the primary airstream mechanism used to produce it. The phonetic characterization of each of these sounds is described in detail in Secs. V A to V D and compared with equivalent sounds attested in human languages, where relevant, to justify the proposed transcription.

A. Articulation of kick/bass drum effects

Three different kick drum effects were demonstrated by the subject, all produced as bilabial ejectives (Figs. 1–3). In all figures showing MR Image sequences, frame numbers are indicated at the bottom left of each image panel. For the video reconstruction rate of 20.8 f.p.s. used in this data, one frame duration is approximately 48 ms.

The effect described as a "punchy kick" (SBN: bf) was produced as a bilabial affricate ejective /p͡f'ːˀ/. Six image frames acquired over a 550 ms interval during the production of one token are shown in Fig. 1. Laryngeal lowering and lingual retraction commence approximately 350 ms before the acoustic release burst; labial approximation commences 230 ms before the burst. Velic raising to seal the nasopharynx off from the oral vocal tract can be observed as the larynx is lowered and the lips achieve closure (frame 97). Glottal closure is clearly evident after the larynx achieves the lowest point of its trajectory (frame 98). Rapid upward movement of the larynx can be observed after glottal adduction, accompanied by rapid raising of the tongue dorsum, resulting in motion blurring throughout the posterior oral and supralaryngeal regions (frame 100).

Mean upward vertical displacement of the glottis during ejective production, measured over five repetitions of the punchykick drum effect, was 21.0 mm. The glottis remained adducted throughout the production of the ejective (frame 101), and was reopened approximately 160 ms after the beginning of the acoustic release burst. At the completion of the ejective, the tongue remained in a low central position (frame 103) resembling the articulatory posture observed during the subject's production of the vowel [ʌ].[2]

In addition to the punchy kick, the subject controlled two variant bass drum effects (SBN: b), both produced as

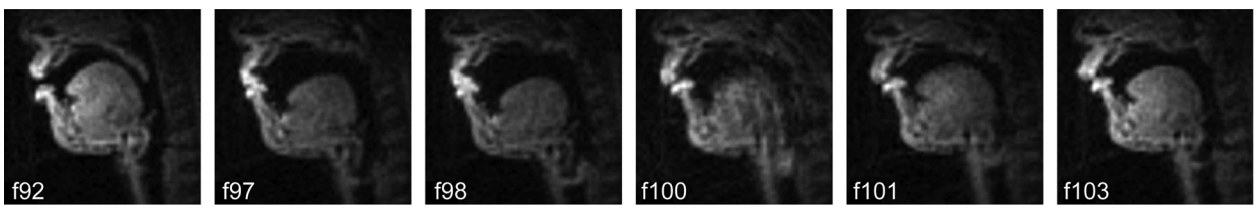

FIG. 1. Articulation of a "punchy" kick drum effect as an affricated labial ejective [p͡f'ːʌ̰]. Frame 92: starting posture; f97: lingual lowering, velic closure; f98: fully lowered larynx, glottalic closure; f100: rapid laryngeal raising accompanied by lingual raising; f101: glottis remains closed during laryngeal raising; f103: glottal abduction; final lingual posture remains lowered.

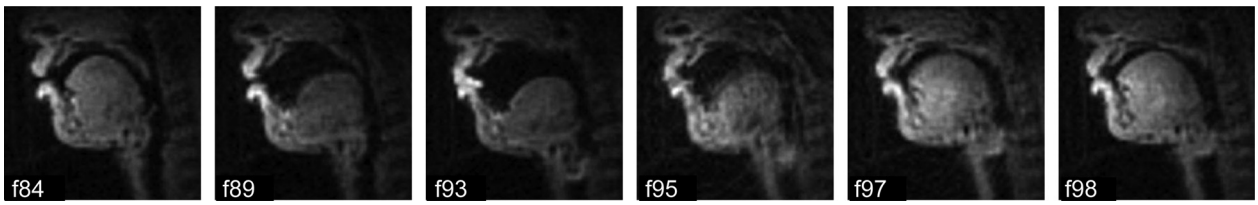

FIG. 2. Articulation of a "thud" kick drum effect as an bilabial ejective [p'ɪ̟]. Frame 84: starting posture; f89: glottal lowering, lingual retraction; f93: fully lowered larynx, sealing of glottalic, velic and labial ports; f95: rapid laryngeal raising accompanied by lingual raising; f97: glottis remains closed during laryngeal raising and lingual advancement; f98: final lingual posture raised and advanced.

unaffricated bilabial ejective stops: a "thud kick," and an "808 kick." Image sequences acquired during production of these effects are shown in Figs. 2 and 3, respectively. The data reveal that although the same basic articulatory sequencing is used, there are minor differences in labial, glottal, and lingual articulation which distinguish each kick drum effect.

In both thud and 808 kick effects, the lips can be seen to form a bilabial seal (Fig. 2, frames 93–95; Fig. 3, frames 80–82), while in the production of the affricated punchy effect, the closure is better characterized as labio-dental (Fig. 1, frames 98–103). Mean upward vertical displacement of the glottis during ejective production, measured over six repetitions of the thud kick drum effect, was 18.6 mm, and in five of the six tokens demonstrated, no glottal abduction was observed after completion of the ejective. Vertical glottal displacement averaged over five tokens of the 808 kick drum effect, was 17.4 mm. Mean duration (oral to glottal release) of the 808 effect was 152 ms.

A final important difference between the three types of kick drum effects produced by this subject concerns lingual articulation. Different amounts of lingual retraction can be observed during laryngeal lowering before production of each ejective. Comparison of the end frames of each image sequence reveals that each effect is produced with a different final lingual posture. These differences can be captured in close phonetic transcription by using unvoiced vowels to represent the final posture of each effect: [p͡f'ːʌ̟](punchy), [p'ɪ̟](thud), and [p'ʊ̟] (808).

These data suggest that the kick drum effects produced by this artist are best characterized as "stiff" (rather than "slack") ejectives, according to the typological classification developed by Lindau (1984), Wright *et al.* (2002), and Kingston (2005): all three effects are produced with a very long voice onset time (VOT), and a highly transient, high amplitude aspiration burst. The durations of these sound effects (152 to 160 ms) are longer than the durations reported for glottalic egressive stops in Tlingit (Maddieson *et al.*, 2001) and Witsuwit'en (Wright *et al.*, 2002), but resemble average release durations of some other Athabaskan glottalic consonants (Hogan, 1976; McDonough and Wood, 2008). In general, it appears that the patterns of coordination between glottal and oral closures in these effects more closely resemble those observed in North American languages, as opposed to African languages like Hausa (Lindau, 1984), where "the oral and glottal closures in an ejective stop are released very close together in time" (Maddieson *et al.*, 2001).

B. Articulation of rim shot effects

Four different percussion effects classified as snare drum "rim shots" were demonstrated by the subject (Table I). Two effects were realized as dorsal stops, differentiated by their airstream mechanisms. Two other rim shot sounds were produced as lingual ingressive consonants, or clicks.

The effect described as "rim shot K" was produced as a voiceless pulmonic egressive dorsal stop, similar to English /k/, but with an exaggerated, prolonged aspiration burst: [k^{h}hː]. Mean duration of the aspiration burst (interval over which aspiration noise exceeded 10% of maximum stop intensity), calculated across three tokens of this effect, was 576 ms, compared to mean VOT durations of 80 ms and 60 ms for voiceless (initial) dorsal stops in American (Lisker and Abramson, 1964) and Canadian English (Sundara, 2005), respectively.

A second effect produced at the same place of articulation was realized as an ejective stop [k'], illustrated in Fig. 4—an image sequence acquired over a 480 ms interval during the production of the second token. Dorsal closure (frame 80) occurs well before laryngeal lowering commences (frame 83). Upward movement of the closed glottis can be observed after the velum closes off the nasopharyngeal port, and glottal closure is maintained until after the dorsal constriction is released (frame 90).

Unlike in the labial kick drum effects, where laryngeal raising was accompanied by rapid movement of the tongue (Figs. 1–3), no extensive lingual movement was observed

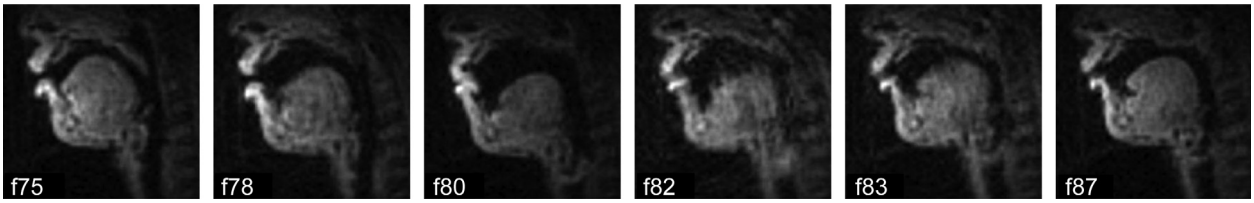

FIG. 3. Articulation of an "808" kick drum effect as an bilabial ejective [p'ʊ̟]. Frame 75: starting posture; f78: lingual lowering, velic closure; f80: fully lowered larynx, glottalic and labial closure; f82: rapid laryngeal raising, with tongue remaining retracted; f83: glottis remains closed during laryngeal raising; f87: glottal abduction; final lingual posture midhigh and back.

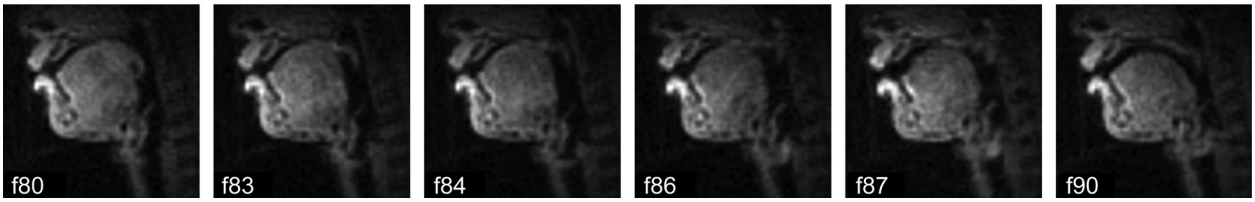

FIG. 4. Articulation of a rim shot effect as a dorsal ejective [k']. Frame 80: dorsal closure; f83: laryngeal lowering, velic raising; f84: velic closure, larynx fully lowered; f86: glottal closure; f87: rapid laryngeal raising; f90: glottis remains closed through completion of ejective and release of dorsal constriction.

during dorsal ejective production in any of the rim shot tokens (frames 86–87). Mean vertical laryngeal displacement, averaged over five tokens, was 14.5 mm. Mean ejective duration (lingual to glottal release) was 142 ms: slightly shorter than, but broadly consistent with, the labial ejective effects described above.

Articulation of the effect described as a "side K rim shot" is illustrated in the image sequence shown in Fig. 5, acquired over a 480 ms interval during the fifth repetition of this effect. The data show that a lingual seal is created between the alveolar ridge and the back of the soft palate (frames 286–290), and that the velum remains lowered throughout. Frames 290–291 reveal that rarefaction and cavity formation occur in the midpalatal region while anterior and posterior lingual seals are maintained, suggesting that the consonantal influx is lateralized, consistent with the subject's description of the click as being produced at "the side of the mouth." The same pattern of articulation was observed in all seven tokens produced by the subject.

Without being able to see inside the cavity formed between the tongue and the roof of the mouth, it is difficult to locate the posterior constriction in these sounds precisely. X-ray data from Traill (1985), for example, reported in Ladefoged and Maddieson (1996), show that back of the tongue maintains a very similar posture across all five types of click in !Xoõ, despite the fact that the lingual cavity varies considerably in size and location. Nevertheless, both lingual posture and patterns of release in this sound effect appear to be consistent with the descriptions of lateral clicks in !Xoõ, N|uu (Miller *et al.*, 2009) and Nama (Ladefoged and Traill, 1984). In summary, this effect appears to be best described as a voiceless uvular nasal lateral click: [ɴ̥|].

The final rim shot effect in the repertoire was described by the subject as "sucking in." The images in Fig. 6 were acquired over a 440 ms interval during the production of the first token of this effect. Like the lateral rim shot, a lingual seal is created in the palatal region with the anterior closure at the alveolar ridge and the posterior closure spread over a broad region of the soft palate (frames 17–20). Once again, the velum remains lowered throughout. The same pattern of articulation was observed in all eight repetitions of this effect. As with the lateral click, we cannot determine exactly where the lingual cavity is formed in this sound effect, nor precisely where and when it is released. Nevertheless, the patterns of tongue movement in these data are consistent with the descriptions of alveolar clicks in !Xoõ, N|uu, and Nama, as well as in Khoekhoe (Miller *et al.*, 2007), so this effect appears to be best described as a voiceless uvular nasal alveolar click: [ɴ̥!].

C. Articulation of snare drum effects

Three different snare drum effects were demonstrated by the subject—a "clap," "meshed," and "no meshed" snare—each produced with different articulatory and airstream mechanisms, described in detail below.

Articulation of the effect described as a "clap snare" is illustrated in the image sequence shown in Fig. 7, acquired over a 240 ms interval during the sixth repetition of this effect. As in the rim shot clicks, a lingual seal is first created along the hard and soft palates, and the velum remains lowered throughout. However, in this case the anterior lingual seal is more anterior (frame 393) than was observed in the lateral and alveolar clicks, the point of influx occurs closer to the subject's teeth (frames 394–395), and the tongue dorsum remains raised higher against the uvular during coronal release. Labial approximation precedes click formation and the labial closure is released with the click. The same pattern of articulation was observed in all six tokens demonstrated by the subject, consistent with the classification of this sound effect as a labialized voiceless uvular nasal dental click: [ɴ̥|ʷ].

The "no mesh" snare drum effect was produced as a labial affricate ejective, similar to the punchy kick drum effect but with a higher target lingual posture: [p͡fʼːʜ̟̊]. The final

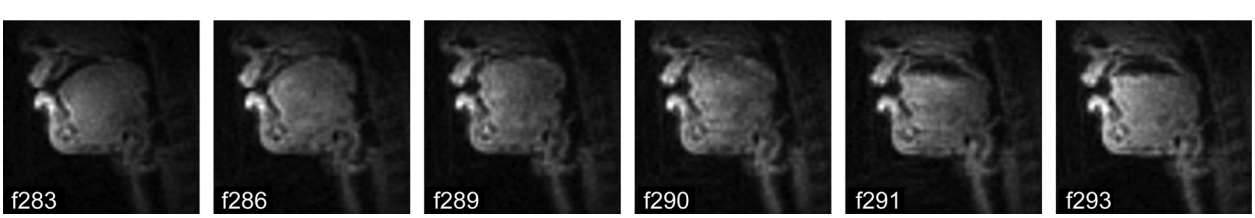

FIG. 5. Articulation of a "side K" rim shot effect as a lateral click [ɴ̥||]. Frame 283: starting posture; f286: lingual raising and advancement towards palate; f289: completion of lingual seal between alveolar ridge and soft palate; f290: beginning of lingual retraction to initiate rarefaction of palatal cavity; f291: lateral influx produced by lowering of tongue body while retaining anterior and posterior lingual seals; f293: final lingual posture. Note that the velum remains lowered throughout click production.

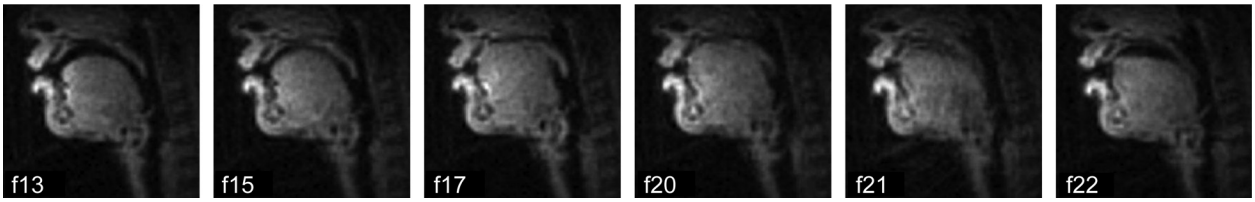

FIG. 6. Articulation of a rim shot effect as an alveolar click [ɴ!]. Frame 13: starting posture; f15: lingual raising and advancement towards palate; f17: completion of lingual seal between alveolar ridge and soft palate; f20–22: rarefaction of palatal cavity; f22: final lingual posture after alveolar release. Note that the velum remains lowered throughout click production.

snare effect, described as "meshed or verby," was produced as a rapid sequence of a dorsal stop followed by a long palatal fricative [kçː]. A pulmonic egressive airstream mechanism was used for all six tokens of the meshed snare effect, but with considerable variability in the accompanying laryngeal setting. In two tokens, complete glottal closure was observed immediately preceding the initial stop burst, and a lesser degree of glottal constriction was observed in another two tokens. Upward vertical laryngeal displacement (7.6 mm) was observed in one token produced with a fully constricted glottis, one token produced with a partially constricted glottis (5.2 mm) and in another produced with an open glottis (11.1 mm). These results suggest that, although canonically pulmonic, the meshed snare effect was variably produced as partially ejective ([k'çː]), or pre-glottalized ([ˀkçː]).

D. Articulation of hi-hat and cymbal effects

Five different effects categorized as "hi-hats" and two effects categorized as cymbals were demonstrated by the subject. All these sounds were produced either as affricates, or as rapid sequences of stops and fricatives articulated at different places.

Articulation of an "open K" hi-hat (SBN: kss) is illustrated in the sequence in Fig. 8, acquired over a 280 ms interval during the fourth repetition. The rapid sequencing of a dorsal stop followed by a long coronal fricative was similar to that observed in the "meshed" snare (Sec. V C), except that the concluding fricative was realized as an apical alveolar sibilant, in contrast to the bunched lingual posture of the palatal sibilant in the snare effect. All seven tokens of this hi-hat effect were primarily realized as pulmonic egressives, again with variable laryngeal setting. Some degree of glottal constriction was observed in five of seven tokens, along with a small amount of laryngeal raising (mean vertical displacement, all tokens = 4.4 mm). The data suggest that the open K hi-hat effect can be characterized as a (partially ejective) pulmonic egressive voiceless stop-fricative sequence [k$^{(\cdot)}$sː].

Two hi-hat effects, the "open T" (SBN: tss) and "closed T" (SBN: t), were realized as alveolar affricates, largely differentiated by their temporal properties. The MRI data show that both effects were articulated as laminal alveolar stops with affricated releases. The closed T effect was produced as a short affricate truncated with a homorganic unreleased stop [t͡s̚], in which the tongue retained a bunched posture throughout. Mean affricate duration was 94 ms (initial stop to final stop, calculated over five tokens). Broadband energy of the short fricative burst extended from 1600 Hz up to the Nyquist frequency (9950 Hz), with peaks at 3794 Hz and 4937 Hz.

The open T effect [t͡sː] was realized without the concluding stop gesture and prolongation of the alveolar sibilant, during which the tongue dorsum was raised and the tongue tip assumed a more apical posture at the alveolar ridge. Mean duration was 410 ms (initial stop to 10% threshold of maximum fricative energy, calculated over five tokens). Broadband energy throughout the fricative phase was concentrated above 1600 Hz, and extended up to the Nyquist frequency (9950 Hz), with peaks at 4883 Hz and 8289 Hz.

Articulation of the hi-hat effect described as "closed: kiss teeth" is illustrated in Fig. 9. The image sequence was acquired over a 430 ms interval during the second of six repetitions of this effect. An elongated constriction was first formed against the alveolar ridge, extending from the back of the upper teeth through to the hard palate (frame 98). Lingual articulation in this effect very closely resembles that of the clap snare (Figs. 5–7), except that a greater degree of labialization can be observed in some tokens. In all six tokens, the velum remained lowered throughout stop production, and the effect concluded with a transient high-frequency fricative burst corresponding to affrication of the initial stop. In all tokens, laryngeal lowering was observed during initial stop production, beginning at the onset of the stop burst, and

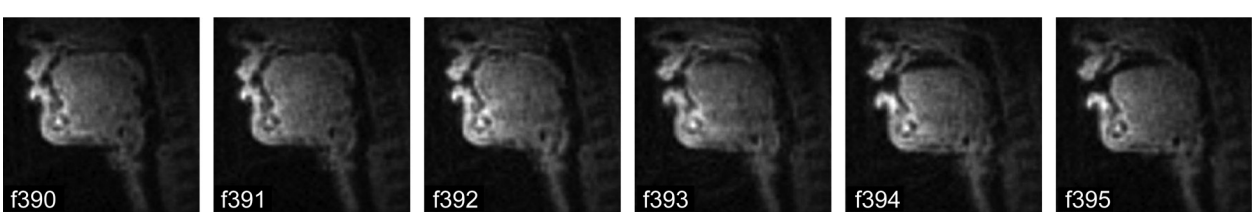

FIG. 7. Articulation of a "clap" snare drum effect as a labialized dental click [ɴ|^{w}]. Frame 390: tongue pressed into palate; f391–392: initiation of downward lingual motion; f393: rarefaction of palatal cavity; f394–395: dental-alveolar influx resulting from coronal lenition while retaining posterior lingual seal; Note that the velum remains lowered throughout click production.

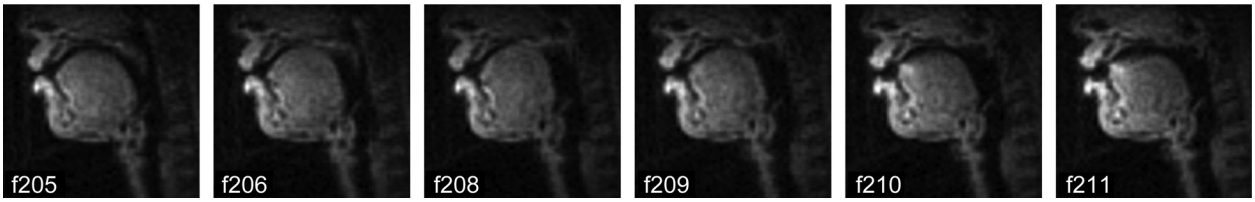

FIG. 8. Articulation of an "open K" hi-hat [ksː]. Frame 205: initial lingual posture; f206–209: dorsal stop production; f209–211: coronal fricative production.

lasting for an average of 137 ms. Mean vertical displacement of the larynx during this period was −3.8 mm. Partial constriction of the glottis during this interval could be observed in four of six tokens. Although this effect was not categorized as a glottalic ingressive, the laryngeal activity suggests some degree of glottalization in some tokens, and is consistent with the observations of Clements (2002), that "larynx lowering is not unique to implosives." In summary, this effect appears to be best described as a pre-labialized, voiceless nasal uvular-dental click [ʷᶯ̥ǀ].

The final hi-hat effect was described as "breathy: in-out." Five tokens were demonstrated, all produced as voiceless fricatives. Mean fricative duration was 552 ms. Broadband energy was distributed up to the nyquist frequency (9900 Hz), with a concentrated noise band located between 1600 and 3700 Hz. Each repetition was articulated with a closed velum, a wide open glottis, labial protrusion, and a narrow constriction formed by an arched tongue dorsum approximating the junction between the hard and soft palates. The effect may be characterized as an elongated labialized pulmonic egressive voiceless velar fricative [xːʷ].

As well as the hi-hat effects described above, the subject demonstrated two cymbal sound effects that he described as "cymbal with a T" and "cymbal with a K." The "T cymbal" was realized as an elongated labialized pulmonic egressive voiceless alveolar-palatal affricate [tɕːʷ]. Mean total duration of five tokens was 522 ms, and broadband energy of the concluding fricative was concentrated between 1700 and 4000 Hz. The "K cymbal" was realized as a pulmonic egressive sequence of a labialized voiceless velar stop followed by a partially labialized palatal fricative [kʷɕːʷ]. Mean total duration of five tokens was 575 ms. Fricative energy was concentrated between 1400 and 4000 Hz.

E. Production of beatboxing sequences

In addition to producing the individual percussion sound effects described above, the subject demonstrated a number of short beatboxing sequences in which he combined different effects to produce rhythmic motifs or "grooves." Four different grooves were demonstrated, each performed at three different target tempi nominated by the subject: slow (∼88 b.p.m.), medium (∼95 b.p.m.), and fast (∼104 b.p.m.). Each groove was realized as a one-, two-, or four-bar repeating motif constructed in a common time signature (4 beat measures), demonstrated by repeating the sequence at least three times. In the last two grooves, the subject improvised on the basic rhythmic structure, adding ornamentation and varying the initial sequence to some extent. Between two and five different percussion elements were combined into each groove (Table II). Broad phonetic descriptions have been used to describe the effects used, as the precise realization of each sound varied with context, tempo and complexity.

VI. TOWARDS A UNIFIED FORMAL DESCRIPTION OF BEATBOXING PERFORMANCE

Having described the elemental combinatorial sound effects of a beatboxing repertoire, we can consider formalisms for describing the ways in which these components are combined in beatboxing performance. Any such representation needs to be able to describe both the musical and linguistic properties of this style—capturing both the metrical structure of the performance and phonetic details of the constituent sounds. By incorporating IPA into standard percussion notation, we are able to describe both these dimensions and the way they are coordinated.

Although practices for representing non-pitched percussion vary (Smith, 2005), notation on a conventional staff typically makes use of a neutral or percussion clef, on which each "pitch" represents an individual instrument in the percussion ensemble. Filled note heads are typically used to represent drums, and cross-headed notes to annotate cymbals; instruments are typically labeled at the beginning of the score or the first time that they are introduced, along with any notes about performance technique (Weinberg, 1998).

The notation system commonly used for music to be performed on a "5-drum" percussion kit (Stone, 1980) is ideal for describing human beatboxing performance because

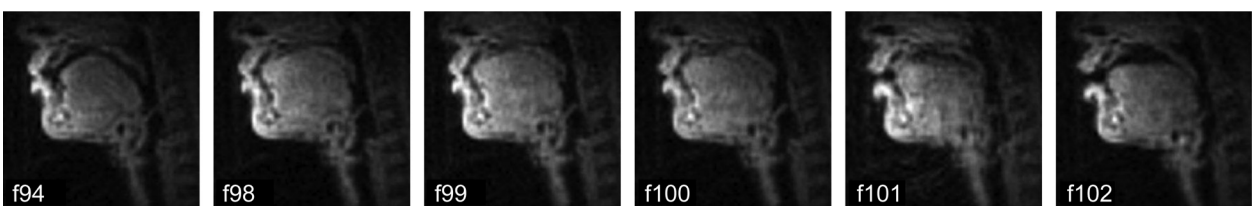

FIG. 9. Articulation of an "closed kiss" hi-hat effect [ʷᶯ̥ǀ]. Frame 94: initial lingual posture; f98: constriction formed against teeth, alveolar ridge and hard palate; f99–101: partial glottal constriction, lowering of tongue and larynx; f102: final lingual posture.

TABLE II. Metrical structure and phonetic composition of four beatboxing sequences (grooves) demonstrated by the subject.

Title	Meter	Bars	Percussion Elements
"Audio 2"	4/4	1	/p'/, /xː/
"Tried by Twelve"	4/4	2	/p'/, /p͡f'/, /ts/
"Come Clean"	4/4	4	/p'/, /p͡f'/, /ts/, /ŋ|/
"Saturday"	4/4	4	/p'/, /p͡f'/, /ts/, /ŋ|/, /ʁ!/

the sound effects in the beatboxer's repertoire typically correspond to similar percussion instruments. The description can be refined and enhanced through the addition of IPA "lyrics" on each note, to provide a more comprehensive description of the mechanisms of production of each sound effect.

For example, the first groove demonstrated by the subject in this experiment, entitled "Audio 2," can be described using the score illustrated in Fig. 10. As in standard non-pitched percussion notation, each instrumental effect—in this case a kick drum and a hi-hat—is represented on a dedicated line of the stave. The specific realization of each percussive element is further described on the accompanying lyrical scores using IPA. Either broad "phonemic" (Fig. 10) or fine phonetic (Fig. 11) transcription of the mechanisms of sound production can be employed in this system.

VII. COMPANION MULTIMEDIA CORPUS

Video and audio recordings of each of the effects and beatboxing sequences described above have been made available online at http://sail.usc.edu/span/beatboxing. For each effect in the subject's repertoire, audio-synchronized video of the complete MRI acquisition is first presented, along with a one-third speed video excerpt demonstrating a single-token production of each target sound effect, and the acoustic signal extracted from the corresponding segment of the companion audio recording. A sequence of cropped, numbered video frames showing major articulatory landmarks involved in the production of each effect is presented with the multimedia, along with close phonetic transcriptions and frame-by-frame annotations of each sequence.

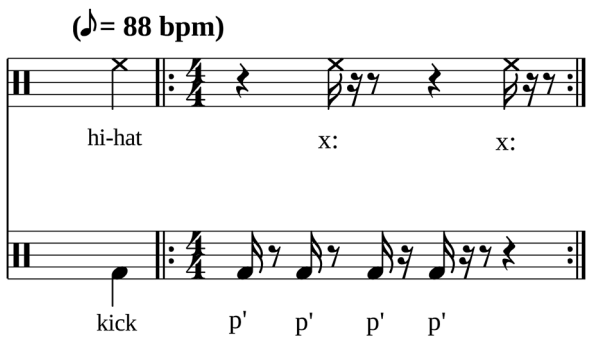

FIG. 10. Broad transcription of beatboxing performance using standard percussion notation: repeated one-bar, two-element groove entitled "Audio 2." Phonetic realization of each percussion element is indicated beneath each voice in the score using broad transcription IPA "lyrics."

VIII. DISCUSSION

The audio and 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 musical, as well as linguistic goals.

A. Phonetic convergence

One of the most important findings of this study is that all of the sounds effects produced by the beatbox artist were able to be described using IPA—an alphabet designed exclusively for the description of contrastive (i.e., meaning encoding) speech sounds. Although this study was limited to a single subject, these data suggest that even when the goals of human sound production are extra-linguistic, speakers will typically marshal patterns of articulatory coordination that are exploited in the phonologies of human languages. To a certain extent, this is not surprising, since speakers of human languages and vocal percussionists are making use of the same vocal apparatus.

The subject of this study is a speaker of American English and Panamanian Spanish, neither of which makes use of non-pulmonic consonants, yet he was able to produce a wide range of non-native consonantal sound effects, including clicks and ejectives. The effects /ŋ||/−/ŋ!/−/ŋ|/ used to emulate the sounds of specific types of snare drums and rim shots appear to be very similar to consonants attested in many African languages, including Xhosa (Bantu language family, spoken in Eastern Cape, South Africa), Khoekhoe (Khoe, Botswana) and !Xóõ (Tuu, Namibia). The ejectives /p'/ and /pf'/ used to emulate kick and snare drums shares the same major phonetic properties as the glottalic egressives used in languages as diverse as Nuxáalk (Salishan, British Columbia), Chechen (Caucasian, Chechnya), and Hausa (Chadic, Nigeria) (Miller *et al.*, 2007; Ladefoged and Maddieson, 1996).

Without phonetic data acquired using the same imaging modality from native speakers, it is unclear how closely non-native, paralinguistic sound effects resemble phonetic equivalents produced by speakers of languages in which these sounds are phonologically exploited. For example, in the initial stages of articulation of all three kick drum effects produced by the subject of this study, extensive lingual lowering is evident (Fig. 1, frame 98; Fig. 2, frame 93; Fig. 3, frame 80), before the tongue and closed larynx are propelled upward together. It would appear that in these cases, the tongue is being used in concert with the larynx to generate a more effective "piston" with which to expel air from the vocal tract.[3] It is not known if speakers of languages with glottalic egressives also recruit the tongue in this way during ejective production, or if coarticulatory and other constraints prohibit such lingual activity.

More typologically diverse and more detailed data will be required to investigate differences in production between these vocal percussion effects and the non-pulmonic

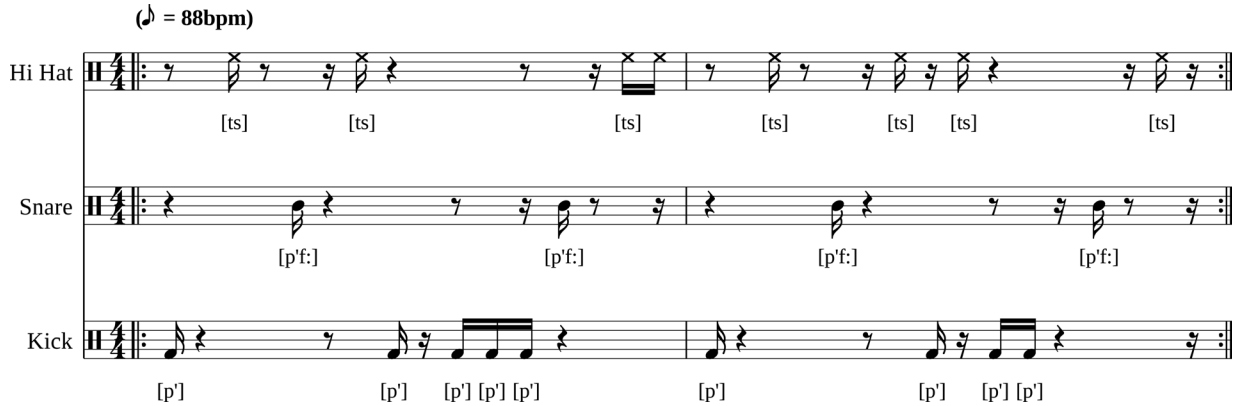

FIG. 11. Fine transcription of beatboxing groove: two-bar, three-element groove entitled "Tried by Twelve" (88 b.p.m.). Detailed mechanisms of production are indicated for each percussion element—"open hat" [ts], "no mesh snare" [p'fː], and "808 kick" [p']—using fine transcription IPA lyrics.

consonants used in different languages. If, as it appears from these data, such differences are minor rather than categorical, then it is remarkable that the patterns of articulatory coordination used in pursuit of paralinguistic goals appear to be consistent with those used in the production of spoken language.

B. Sensitivity to and exploitation of fine phonetic detail

Another important observation to be made from this data is that the subject appears to be highly sensitive to ways in which fine differences in articulation and duration can be exploited for musical effect. Although broad classes of sound effects were all produced with the same basic articulatory mechanisms, subtle differences in production were observed between tokens, consistent with the artist's description of these as variant forms.

For example, a range of different kick and snare drum effects demonstrated in this study were all realized as labial ejectives. Yet the subject appears to have been sensitive to ways that manipulation of the tongue mass can affect factors such as back-cavity resonance and airstream transience, and so was able to control for these factors to produce the subtle but salient differences between the effects realized as [p͡fˈːʌ̹], [pˈɪ̹], [pˈʊ̹], and [p͡fˈːɯ̹].

This musically motivated manipulation of fine phonetic detail—while simultaneously preserving the basic articulatory patterns associated with a particular class of percussion effects—may be compared to the phonetic manifestation of affective variability in speech. In order to convey emotional state and other paralinguistic factors, speakers routinely manipulate voice quality (Scherer, 2003), the glottal source waveform (Gobl and Ní Chasaide, 2003; Bone *et al.*, 2010), and supralaryngeal articulatory setting (Erickson *et al.*, 1998; Nordstrand *et al.*, 2004), without altering the fundamental phonological information encoded in the speech signal. Just as speakers are sensitive to ways that phonetic parameters may be manipulated within the constraints dictated by the underlying sequences of articulatory primitives, the beatbox artist is able to manipulate the production of a percussion element for musical effect within the range of articulatory possibilities for each class of sounds.

C. Goals of production in paralinguistic vocalization

A pervasive issue in the analysis and transcription of vocal percussion is determining which aspects of articulation are pertinent to the description of each sound effect. For example, differences in tongue body posture were observed throughout the production of each of the kick drum sound effects—both before initiation of the glottalic airstream and after release of the ejective (Sec. V A). It is unclear which of these tongue body movements are primarily related to the mechanics of production—in particular, airstream initiation—and which dorsal activity is primarily motivated by sound shaping.

Especially in the case of vocal percussion effects articulated primarily as labials and coronals, we would expect to see some degree of independence between tongue body/root activity and other articulators, much as vocalic coarticulatory effects are observed to be pervasive throughout the production of consonants (Wood, 1982; Gafos, 1999). In the vocal percussion repertoire examined in this study, it appears that tongue body positioning *after* consonantal release is the most salient factor in sound shaping: the subject manipulates target dorsal posture to differentiate sounds and extend his repertoire. Vocalic elements are included in the transcriptions in Table I only when the data suggest that tongue posture is actively and contrastively controlled by the subject. More phonetic data is needed to determine how speakers control post-ejective tongue body posture, and the degree to which the tongue root and larynx are coupled during the production of glottalic ejectives.

D. Compositionality in vocal production

Although beatboxing is fundamentally an artistic activity, motivated by musical, rather than linguistic instincts, sound production in this domain—like phonologically motivated vocalization—exhibits many of the properties of a discrete combinatorial system. Although highly complex sequences of articulation are observed in the repertoire of the beatboxer, all of the activity analyzed here is ultimately reducible to coordinative structures of a small set of primitives involving pulmonic, glottal, velic and labial states, and the lingual manipulation of stricture in different regions of the vocal tract.

Further examination of beatboxing and other vocal imitation data may shed further light on the nature of compositionality in vocal production—the extent to which the generative primitives used in paralinguistic tasks are segmental, organic or gestural in nature, and whether these units are coordinated using the same principles of temporal and spatial organization which have been demonstrated in speech production (e.g., Saltzman and Munhall, 1989).

E. Relationships between production and perception

Stowell and Plumbley (2010, p. 2) observe that "the musical sounds which beatboxers imitate may not sound much like conventional vocal utterances. Therefore the vowel-consonant alternation which is typical of most use of voice may not be entirely suitable for producing a close auditory match." Based on this observation, they conclude that "beatboxers learn to produce sounds to match the sound patterns they aim to replicate, attempting to overcome linguistic patternings. Since human listeners are known to use linguistic sound patterns as one cue to understanding a spoken voice… it seems likely that avoiding such patterns may help maintain the illusion of non-voice sound." The results of this study suggest that, even if the use of non-linguistic articulation *is* a goal of production in human beatboxing, artists may be unable to avoid converging on some patterns of articulation which have been exploited in human languages. The fact that musical constraints dictate that these articulations may be organized suprasegmentally in patterns other than those which would result from syllabic and prosodic organization may contribute to their perception as non-linguistic sounds, especially when further modified by the skillful use of "close-mic" technique.

F. Approaches to beatboxing notation

Describing beatboxing performance using the system outlined in Sec. VI offers some important advantages over other notational systems that have been proposed, such as mixed symbol alphabets (Stowell, 2012), Standard Beatboxing Notation (Splinter and Tyte, 2012) and English-based equivalents (Sinyor *et al.*, 2005), and the use of tablature or plain text (Stowell, 2012) to indicate metrical structure. The system proposed here builds on two formal notation systems with rich traditions, that have been developed, refined, and accepted by international communities of musicians and linguists, and which are also widely known amongst non-specialists.

The integration of IPA and standard percussion notation makes use of established methodologies that are sufficiently rich to describe any sound or musical idea that can be produced by a beatboxer. There are ways of making sounds in the vocal tract that are not represented in the IPA because they are unattested, have marginal status or serve only a special role in human language (Eklund, 2008). Yet because the performer's repertoire makes use of the same vocal apparatus and is limited by the same physiological constraints that have shaped human phonologies, the International Phonetic Alphabet and its extensions provides an ample vocabulary with which to describe the vast majority of sound effects used by (and, we believe, potentially used by) beatboxers.

Standard Beatboxing Notation has the advantage that it uses only Roman orthography, and appears to have gained some currency in the beatboxing community, but it remains far from being standardized and is hampered by a considerable degree of ambiguity. Many different types of kick and bass drum sounds, for example, are all typically transcribed as "b" (see Splinter and Tyte, 2012), and conventions vary as to how to augment the basic SBN vocabulary with more detail about the effects being described. The use of IPA (Stowell, 2012) eliminates all of these problems, allowing the musician, artist, or observer to unambiguously describe any sequence of beatboxing effects at different levels of detail.

The examples illustrated in Figs. 10 and 11 also demonstrate how the musical characteristics of beatboxing performance can be well described using standard percussion notation. In addition, it would be possible to make use of other conventions of musical notation, including breath and pause marks, note ornamentation, accents, staccato, fermata, and dynamic markings to further enrich the utility of this approach as a method of transcribing beatboxing performance. Stone (1980, pp. 205–225) outlines the vast system of extended notation that has been developed to describe the different ensembles, effects and techniques used in traditional percussion performance; many of these same notation conventions could easily be used in the description of human beatboxing performance, where IPA and standard musical notation is not sufficiently comprehensive.

G. Future directions

This work represents a first step towards the formal study of the paralinguistic articulatory phonetics underlying an emerging genre of vocal performance. An obvious limitation of the current study is the use of a single subject. Because beatboxing is a highly individualized artistic form, examination of the repertoires of other beatbox artists would be an important step towards a more comprehensive understanding of the range of effects exploited in beatboxing, and the articulatory mechanisms involved in producing these sounds.

More sophisticated insights into the musical and phonetic characteristics of vocal percussion will emerge from analysis of acoustic recordings along with the companion articulatory data. However, there are obstacles preventing more extensive acoustic analysis of data acquired using current methodologies. The confined space and undamped surfaces within an MRI scanner bore creates a highly resonant, echo-prone recording environment, which also varies with the physical properties of the subject and the acoustic signature of the scan sequence. The need for additional signal processing to attenuate scanner noise (Bresch *et al.*, 2006) further degrades the acoustic fidelity of rtMRI recordings which, while perfectly adequate for the qualitative analysis of human percussion effects presented here, do not permit detailed time-series or spectral analysis. There is a need to develop better in-scanner recording and noise-reduction

technologies for rtMRI experimentation, especially for studies involving highly transient sounds, such as clicks, ejectives, and imitated percussion sounds.

Further insights into the mechanics of human beatboxing will also be gained through technological improvements in MR imaging. The use of imaging planes other than midsagittal will allow for finer examination of many aspects of articulation that may be exploited for acoustic effect, such as tongue lateralization and tongue groove formation. Since many beatbox effects appear to make use of non-pulmonic airstream mechanisms, axial imaging could provide additional detail about the articulation of the larynx and glottis during ejective and implosive production.

Because clicks also carry a high functional load in the repertoire of many beatbox artists, higher-speech imaging of the hard palate region would be particularly useful. One important limitation of the rtMRI sequences used in this study is that, unlike sagittal X-ray (Ladefoged and Traill, 1984), the inside of the cavity is not well resolved during click production; as a result, the precise location of the lingual-velaric seal is not evident. Finer spatial sampling over thinner sagittal planes would provide greater insights into this important aspect of click production. Strategic placement of coronal imaging slices would provide additional phonetic detail about lingual coordination in the mid-oral region. Lateral clicks, which are exploited by many beatbox artists (Tyte, 2012), can only be properly examined using coronal or parasagittal slices, since the critical articulation occurs away from the midsagittal plane. New techniques allowing simultaneous dynamic imaging of multiple planes located at critical regions of the tract (Kim *et al.*, 2012) hold promise as viable methods of investigating these sounds, if temporal resolution can be improved.

Most importantly, there is a need to acquire phonetic data from native speakers of languages whose phonologies include some of the sounds exploited in the beatboxing repertoire. MR images of natively produced ejectives, implosives and clicks—consonants for which there is little non-acoustic phonetic data available—would provide tremendous insights into the articulatory and coordinative mechanisms involved in the generation of these classes of sounds, and the differences between native, non-native, and paralinguistic production.

Highly skilled beatbox artists such as Rahzel are capable of performing in a way which creates the illusion that the artist is simultaneously singing and providing their own percussion accompaniment, or simultaneous beatboxing while humming (Stowell and Plumbley, 2008). Such illusions raise important questions about the relationship between speech production and perception, and the mechanisms of perception that are engaged when a listener is presented with simultaneous speech and music signals. It would be of great interest to study this type of performance using MR Imaging, to examine the ways in which linguistic and paralinguistic gestures can be coordinated.

IX. CONCLUSIONS

Real-Time Magnetic Resonance Imaging has been shown to be a viable method with which to examine the repertoire of a human beatboxer, affording novel insights into the mechanisms of production of the imitation percussion effects that characterize this performance style. The data reveal that beatboxing performance involves the use of many of the airstream mechanisms found in human languages. The study of beatboxing performance has the potential to provide important insights into articulatory coordination in speech production, and mechanisms of perception of simultaneous speech and music.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant R01 DC007124-01. Special thanks to our experimental subject for demonstrating his musical and linguistic talents in the service of science. We are especially grateful to three anonymous reviewers for their extensive comments on earlier versions of this manuscript.

[1]Preliminary analyzes of a subset of this corpus were originally described in Proctor *et al.* (2010b)

[2]All descriptions of dorsal posture were made by comparison to vowels produced by the subject in spoken and sung speech, and in the set of reference vowels elicited using the [h_d] corpus.

[3]Special thanks to an anonymous reviewer for this observation.

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

Bone, D., Kim, S., Lee, S., and Narayanan, S. (**2010**). "A study of intra-speaker and interspeaker affective variability using electroglottograph and inverse filtered glottal waveforms," in *Proc. Interspeech*, Makuhari, pp. 913–916.

Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., and Narayanan, S. (**2008**). "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]," IEEE Signal Process. Mag. **25**, 123–132.

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S. (**2006**). "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791–1794.

Clements, N. (**2002**). "Explosives, implosives, and nonexplosives: The linguistic function of air pressure differences in stops," in *Laboratory Phonology*, edited by C. Gussenhoven and N. Warner (Mouton De Gruyter, Berlin), Vol. 7, pp. 299–350.

Eklund, R. (**2008**). "Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech," J. Int. Phonetic. Assoc. **38**, 235–324.

Erickson, D., Fujimura, O., and Pardo, B. (**1998**). "Articulatory correlates of prosodic control: Emotion and emphasis," Lang. Speech **41**, 399–417.

Gafos, A. (**1999**). *The Articulatory Basis of Locality in Phonology* (Garland, New York), pp. 272.

Gobl, C., and Ní Chasaide, A. (**2003**). "The role of voice quality in communicating emotion, mood and attitude," Speech Comm. **40**, 189–212.

Hess, M. (**2007**). *Icons of Hip Hop: An Encyclopedia of the Movement, Music, and Culture* (Greenwood Press, Westport), pp. 640.

Hogan, J. (**1976**). "An analysis of the temporal features of ejective consonants," Phonetica **33**, 275–284.

Kapur, A., Benning, M., and Tzanetakis, G. (**2004**). "Query-by-beat-boxing: Music retrieval for the DJ," in *Proc. 5th Intl. Conf. on Music Information Retrieval (ISMIR)*, Barcelona, pp. 170–178.

Kim, Y.-C., Proctor, M. I., Narayanan, S. S., and Nayak, K. S. (**2012**). "Improved imaging of lingual articulation using real-time multislice MRI," J. Magn. Resonance Imaging **35**, 943–948.

Kingston, J. (**2005**). "The phonetics of Athabaskan tonogenesis," in *Athabaskan Prosody*, edited by S. Hargus and K. Rice (John Benjamins, Amsterdam), pp. 137–184.

Ladefoged, P., and Maddieson, I. (**1996**). *The Sounds of the World's Languages* (Blackwell, Oxford), pp. 426.

Ladefoged, P., and Traill, A. (**1984**). "Linguistic phonetic descriptions of clicks," Language **60**, 1–20.
Lederer, K. (**2005**). "The phonetics of beatboxing," BA dissertation, Leeds Univ., UK.
Lindau, M. (**1984**). "Phonetic differences in glottalic consonants," J. Phonetics **12**, 147–155.
Lisker, L., and Abramson, A. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.
Maddieson, I., Smith, C., and Bessell, N. (**2001**). "Aspects of the phonetics of Tlingit," Anthropolog. Ling. **43**, 135–176.
McDonough, J., and Wood, V. (**2008**). "The stop contrasts of the Athabaskan languages," J. Phonetics **36**, 427–449.
McLean, A., and Wiggins, G. (**2009**). "Words, movement and timbre," in *Proc. Intl. Conf. on New Interfaces for Musical Expression (NIME'09)*, edited by A. Zahler and R. Dannenberg (Carnegie Mellon Univ., Pittsburgh, PA), pp. 276–279.
Miller, A., Namaseb, L., and Iskarous, K. (**2007**). "Tongue body constriction differences in click types," in *Laboratory Phonology*, edited by J. Cole and J. Hualde (Mouton de Gruyter, Berlin), Vol. 9, 643–656.
Miller, A. L., Brugman, J., Sands, B., Namaseb, L., Exter, M., and Collins, C. (**2009**). "Differences in airstream and posterior place of articulation among Nuu clicks" JIPA **39**, 129–161.
Narayanan, S., Bresch, E., Ghosh, P. K., Goldstein, L., Katsamanis, A., Kim, Y.-C., Lammert, A., Proctor, M. I., Ramanarayanan, V., and Zhu, Y. (**2011**). "A multimodal real-time MRI articulatory corpus for speech research," in *Proc. Interspeech*, Florence, pp. 837–840.
Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (**2004**). "An approach to realtime magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776.
Nordstrand, M., Svanfeldt, G., Granstrm, B., and House, D. (**2004**). "Measurements of articulatory variation in expressive speech for a set of Swedish vowels," Speech Comm. **44**, 187–196.
Proctor, M. I., Bone, D., and Narayanan, S. S. (**2010a**). "Rapid semi-automatic segmentation of real-time Magnetic Resonance Images for parametric vocal tract analysis," in *Proc. Interspeech*, Makuhari, pp. 23–28.
Proctor, M. I., Nayak, K. S., and Narayanan, S. S. (**2010b**). "Linguistic and para-linguistic mechanisms of production in human "beatboxing": A rtMRI study," in *Proc. Intersinging*, Univ. of Tokyo, pp. 1576–1579.
Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. **1**, 333–382.
Scherer, K. (**2003**). "Vocal communication of emotion: A review of research paradigms," Speech Comm. **40**, 227–256.
Sinyor, E., Rebecca, C. M., Mcennis, D., and Fujinaga, I. (**2005**). "Beatbox classification using ACE," in *Proc. Intl. Conf. on Music Information Retrieval*, London, pp. 672–675.
Smith, A. G. (**2005**). "An examination of notation in selected repertoire for multiple percussion," Ph.D. dissertation, Ohio State Univ., Columbus, OH.
Splinter, M., and Tyte, G. (**2006–2012**). "Standard beatbox notation," http://www.humanbeatbox.com/tips/p2_articleid/231 (Last viewed February 16, 2012).
Stone, K. (**1980**). *Music Notation in the Twentieth Century: A Practical Guidebook* (W. W. Norton, New York), 357 pp.
Stowell, D. (**2008–2012**). "The beatbox alphabet," http://www.mcld.co.uk/beatboxalphabet/ (Last viewed February 22, 2012).
Stowell, D. (**2010**). "Making music through real-time voice timbre analysis: machine learning and timbral control," Ph.D. dissertation, School of Electronic Engineering and Computer Science, Queen Mary Univ., London.
Stowell, D., and Plumbley, M. D. (**2008**). "Characteristics of the beatboxing vocal style," Technical Report C4DM-TR-08-01 (Centre for Digital Music, Dep. of Electronic Engineering, Univ. of London, London), pp. 1–4.
Stowell, D., and Plumbley, M. D. (**2010**). "Delayed decision-making in real-time beatbox percussion classification," J. New Music Res. **39**, 203–213.
Sundara, M. (**2005**). "Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French," J. Acoust. Soc. Am. **118**, 1026–1037.
Traill, A. (**1985**). *Phonetic and Phonological Studies of!Xoõ Bushman* (Helmut Buske, Hamburg, Germany), 215 pp.
Tyte, G. (**2012**). "Beatboxing techniques," www.humanbeatbox.com (Last viewed February 16, 2012).
Weinberg, N. (**1998**). *Guide to Standardized Drumset Notation* (Percussive Arts Society, Lawton, OK), pp. 43.
Wood, S. (**1982**). "X-Ray and model studies of vowel articulation," in *Working Papers in Linguistics* (Dep. Linguistics, Lund Univ. Lund, Sweden), Vol. 23, pp. 192.
Wright, R., Hargus, S., and Davis, K. (**2002**). "On the categorization of ejectives: Data from Witsuwit'en" J. Int. Phonetics Assoc. **32**, 43–77.