



181st Meeting of the Acoustical Society of America

Seattle, Washington

29 November - 3 December 2021

*Speech Communication: Paper 3aSC7

In and out: production mechanisms in Human Beatboxing

Alexis Dehais-Underdown

Laboratoire de Phonétique et Phonologie, Université Sorbonne Nouvelle Paris 3, Paris, Ile de France, 75005, FRANCE; alexis.dehais-underdown@sorbonne-nouvelle.fr

Paul Vignes, Lise Crevier-Buchman and Didier Demolin

Laboratoire de Phonétique et Phonologie, Paris, Ile de France, FRANCE; vignes.paul@gmail.com, lise.buchman1@gmail.com; didier.demolin@sorbonne-nouvelle.fr

Human Beatboxing (HBB) is a musical technique produced with vocal tract movements. In this study we investigate HBB production mechanisms of 9 drum imitations produced in isolation by 5 beatboxers. Based on aerodynamic (intraoral pressure and oral airflow), acoustic and laryngoscopic data we were able to identify gestures. Results show beatboxers use a wider range of production mechanisms compared to speech production. Indeed, we found both ingressive and egressive airflow initiated either in the mouth, the larynx or the lungs. For a same sound we found that participants were able to use different mechanisms such as a combination of glottalic and pulmonic airstreams. This raises questions concerning the coordination of initiatory mechanisms and its planification. HBB differs from speech production because it is a musical system that does not rely on linguistic constraints such as intelligibility of semantic meaning so patterns are not as restricted as in speech. Human Beatboxing is a great paradigm to explore the vocal tract capacities during production. Beatboxers have an extended control of their vocal tract to combine different initiation and articulation mechanisms. HBB offers new perspectives on articulatory complexity, phonetic diversity of the world's languages and potential application for speech pathology.

***2nd Prize Best Poster Presentation by the Speech Communication Technical Committee**

1. INTRODUCTION

A. BEATBOXING AND SPEAKING

Human Beatboxing (HBB) is a musical technique that uses the vocal tract to imitate musical instruments. It is not the only vocal technique using non-linguistic sounds, we could refer, for example, to Ella Fitzgerald and her scatting talent. Similar to speech, HBB relies on the selection and combination of smaller units into larger ones. However, unlike linguistic systems, HBB has no meaning: while we speak to be understood, beatboxers do not perform to be understood. Speech production obeys to linguistic constraints to ensure effective communication by using a finite set of units and combinations. Such constraints do not apply to Beatboxing since its units are meaningless and thus the system is not defined by a set of finite gestures and combinations.

HBB is a peculiar technique characterized, mostly, by (1) articulatory precision and (2) breathing control. Concerning articulation, beatboxers can produce a larger number of production mechanisms compared to a native speaker of a given language. Methods for studying beatboxing articulation are MRI (Proctor et al., 2013; Blaylock et al., 2017), EMA (Paroni et al., 2021) and fibroscopy (De Torcy et al., 2014; Sapthavee et al., 2014; Dehais-Underdown et al., 2019). MRI and EMA studies describe the diversity of beatboxing articulation for producing beatboxing sounds. Fibroscopic studies describe the laryngeal articulator and showed beatboxers do not present any signs of injuries of the laryngeal structures. MRI and EMA studies also discuss the nature and the function of beatboxing sounds and their relationship with speech units. Paroni et al. (2021) define sounds as “boxemes by analogy with speech phonemes”, the authors do not define the nature and if, by analogy to speech phonemes, “boxemes” function is contrastive and distinctive. The nature of HBB units are still under investigation, further investigation may give more information about their function. For the time, it seems more cautious not to define HBB units by analogy to speech units because of the absence of semantic meaning. Concerning breathing control, the most striking aspect is the fact that beatboxers have an uninterrupted vocal production without running out of air. In fact, they produce ingressive sounds to insure a continuous production (Stowell and Plumley, 2008). HBB studies on aerodynamics and ventilation are few (Dehais-Underdown et al., 2019; Paroni et al., 2021). Paroni et al. (2021) used EMA and respiratory inductance plethysmography to infer non-pulmonic airflow while Dehais-Underdown et al. (2019) used direct recordings of airflow and intraoral pressure. More studies are needed to understand the reorganization and the coordination of articulation and breathing.

Our research is interested in the human vocal tract capacities and limits. The knowledge on the vocal tract capacities and its functioning mostly comes from speech research. However, HBB questions our understanding of the vocal tract capacities. Indeed, beatboxers are able to produce a large variety of sounds that are not attested in any phonological systems (Blaylock et al., 2017). Since speech and music have different production goals and depend on different systemic constraints (e.g. linguistic constraints *versus* aesthetic constraints), the use of the vocal tract has to be different too. We hypothesize that beatboxers acquire a more accurate and extended control on aeromechanical constraints of the vocal tract allowing them to use a larger number of production mechanisms.

B. AEROMECHANICS OF SPEECH

Aeromechanics is a branch of aerodynamics that deals with the motion of gases. It is relevant for speech science because sounds are the result of air pressure variations in the vocal tract (Hixon et al., 2018). In this study we will focus on 2 aerodynamic principles, *Boyle's Law* and the *Equalization of Pressure* (Catford et al., 1977; Gick et al., 2012).

Boyle's Law stipulates that, in a close system with a constant temperature, pressure and volume are in an inverse relationship. Transposed to speech production, when the volume of the vocal tract is diminishing, pressure rises up and vice-versa. Catford defines pressure limits in the vocal tract between -100hPa and $+160\text{hPa}$, though a more reduced range is actually observed in speech production. Linguistic data on pressure in the oro-pharyngeal cavity ranges from 3 to 15hPa and up to 40-50hPa for languages having ejectives in their phonological inventory.

Pressure Equalization specifies that air moves from high pressure regions to low pressure ones (Gick et al., 2012). The result of equalization is air movement in a direction; we call it *Directionality* of the airflow. Transposed to speech production, when pressure is lower in front the constriction point, air will move to this low pressure region. There are 2 airflow directions: egressive flow and ingressive flow. Egressive flow is very common since speech is produced during the exhalation phase of breathing. Specific sounds are phonologically ingressive (i.e. clicks and implosives). Words or small utterances can be produced with a pulmonic ingressive airflow (Eklund, 2008). However, “ingressiveness” is not a relevant linguistic feature in speech production. Concerning the transcription of egressive and ingressive flow, we

chose to “mark” ingressive flow with a downward arrow [↓] and the absence of an arrow means the flow is egressive. Finally, there are 3 locations where airflow originates: the lungs (i.e. pulmonic airflow), the larynx (i.e. glottalic airflow) or the mouth (i.e. buccal airflow). We summarized these parameters in Table 1 (note that we will use the phonetic description column for describing beatboxed sounds and that we prefer to use the term buccal instead of velaric to refer to the airstream initiation mechanism).

Table 1: Initiation Parameters

Location	Direction	Phonetic Description	Beatbox Example
Lungs	Egressive	Pulmonic Egressive	Cough Snare [f̄h]
	Ingressive	Pulmonic Ingressive	Inward K-Snare [↓kl̄]
Larynx	Egressive	Glottalic Egressive	Classic Kick [p̄']
	Ingressive	Glottalic Ingressive	Inward Classic Kick [↓p]
Mouth	Egressive	Buccal Egressive	Humming Hi-Hat {ts̄}
	Ingressive	Buccal Ingressive	Humming K-Snare {↓kl̄}

In this paper we will refer to 3 parameters: *volume*, *pressure* and *volume velocity*. The *volume* refers to the space that air fills in every dimension. We express the volume in *cubic decimeters* (dm^3). During speech production, volume continuously change and generates pressure variations. The *pressure* is the force applied on a surface in an uniform and perpendicular way. We express the pressure in *hectopascal* (hPa). The *volume velocity* is the displacement of an air volume per unit of time. Airflow velocity depends on the constriction degree; velocity is higher when passing through a narrow constriction. Volume velocity is expressed in *cubic decimeter per second* (dm^3/s). Catford (1977) defines an average around 0.10-0.25 dm^3/s with peaks of $1dm^3/s$ for [h].

C. THE LARYNGEAL ARTICULATOR MODEL

The *Laryngeal Articulator Model or LAM* (Esling et al., 2019) seeks to change the simplistic view that reduces the larynx to vocal folds vibration. Instead, the LAM proposes a revision of articulatory and acoustic mechanisms of the lower vocal tract during speech production. The model puts laryngeal articulations on the same level as supralaryngeal activity for a better understanding of voice quality features.

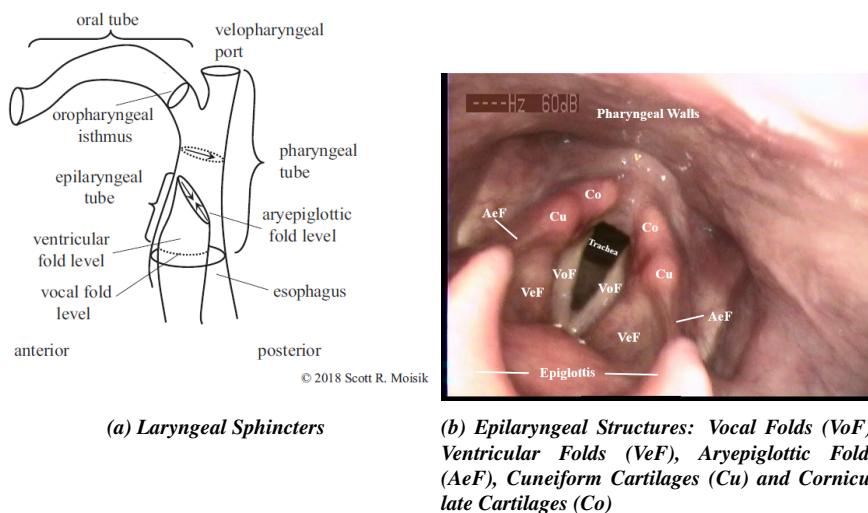


Figure 1: Anatomical Structures of the Laryngeal Articulator

The laryngeal articulator is composed of different parts (Figure 1a): pharyngeal tube (i.e. lower pharynx and tongue root), epilaryngeal tube (i.e. Aryepiglottic Folds and epiglottis) and the glottal level (i.e. Vocal Folds and

Ventricular Folds). Folds are the most efficient structures to produce vibratory mechanism. Three types of folds are located in the larynx (Figure 1b), (1) Vocal Folds (VoF), (2) Ventricular Folds (VeF) and (3) Aryepiglottic Folds (AeF). VoF are the primary source of vibration during voice production. Coupling of VeF and VoF can occur and modify the vibratory pattern by changing the vibrating mass. Finally, AeF can also vibrate. Contrary to VoF and VeF, when an anteroposterior compression of AeF is achieved Corniculate Cartilages are brought on top of the epiglottis tubercle and constrict the epilaryngeal tube. AeF do not vibrate on one another but on the root of the epiglottis.



Figure 2: LAM Revised Open-Closed Continuum (Esling et al., 2019)

By redefining the laryngeal articulator role in speech production, the authors also propose a redefinition of the *Open-Close Continuum* (Gordon and Ladefoged, 2001). Originally, the continuum was referring to glottal settings for the production of phonation types. It ranged from voiceless configuration with a maximal aperture to a glottal stop. Figure 2 illustrates the revision of the open-closed continuum by rearranging laryngeal states within a multiplanar approach; the revised continuum is now described at different 3 levels. The first level is VoF complete closure achieved on configuration 5 (Figure 2). The second is VeF complete closure (Figure 2 configuration 6). The third is Aryepiglottic closure (Figure 2 configuration 7) when the closure is achieved, Cuneiform Tubercles are being pulled on the epiglottis.

Constrictions of the laryngeal articulator can modify the pulmonic airflow and generate voicing, frication, trilling or simply stop the airflow. Moreover, the larynx can initiate egressive or ingressive airflows with a *glottalic mechanism*. When the glottis is closed, or adducted, vertical movements of the larynx and volume expansion (i.e. implosives) or reduction (i.e. ejectives) may create sufficient pressure changes to set air in motion.

2. METHODS

In order to investigate HBB production mechanisms and to gain new insights on vocal tract capacities, we designed an experiment based on aerodynamic, laryngoscopic and acoustic data. The data presented in this study was extracted from a larger corpus of physiological data on HBB. Both the protocol and the corpus were created in collaboration with a professional beatboxer. Each participant gave his or her written consent for this experiment and the agreement from the ethical committee of the hospital.

A. CORPUS

The corpus is composed of controlled and freestyle productions. Here, we will focus on the controlled productions of sounds in isolation. The phonetic structure of the extracted sounds is diverse because they contain a wide range of production mechanisms such as buccal articulations, ingressive airflows, trilling or glottalic airflows. Table 2 presents the beatboxed sounds extracted from the database. Similar to Proctor et al. (2013), we chose to adapt IPA symbols (International Phonetic Association, 1999) and their extensions for disordered speech (Ball et al., 2018; Bernhardt and Ball, 1993) to transcribe beatboxed sounds. Arrows indicate airflow directions; sounds transcribed within brackets (e.g. {p}) indicate buccal articulations, that is, sound produced with the same lingual mechanism as clicks. Buccal articulations in HBB are used for humming. Humming is a strategy that decouples the oral and laryngo-pharyngeal cavities in order to produce articulations and voice simultaneously.

B. PROTOCOL

We designed a production task based on the repetition of audio recordings of sounds that were acquired beforehand. The experiment took place at Hôpital Foch (Suresnes, France). One session was dedicated to laryngoscopic data acquisition and a second session was dedicated to aerodynamic and acoustic recordings. The participants were asked

Table 2: Corpus of beatboxed sounds.

Sound	Beatboxing Category	Transcription	Description
Classic Kick		[p']	glottalic egressive bilabial stop
Humming Kick		{p}	buccal egressive bilabial stop
Closed Hi-Hat		[ts']	glottalic egressive coronal affricate
Humming Hi-Hat		{ts}	buccal egressive coronal affricate
Inward K-Snare		[↓KL]	pulmonic ingressive lateral velar affricate
Humming K-Snare		{↓KL}	buccal ingressive lateral velar affricate
Cough Snare		[?h]	pulmonic egressive epilaryngeal affricate
Lips Roll		[↓B ¹]	voiceless pulmonic ingressive lateral bilabial trill
Humming Lips Roll		{↓B ¹ }	voiceless buccal ingressive lateral bilabial trill

to listen and repeat the beatboxing sounds and patterns and they were allowed to ask questions about them. Each sound was produced in isolation 8 times. Five professional participants were recorded: 1 pilot subject (VP), 2 female beatboxers (LJ and IA) and 2 male beatboxers (CJ and GA).

i. Laryngoscopy

Laryngoscopic data was acquired by the team's MD with a Xion video-stroboscopic system using a flexible fiberscope with a rate of 25 frames per second. The acoustic signal was acquired through the fiberscope microphone. A 2% Xylocaine anesthesia was administered before the endoscope was inserted through the nose into the pharyngeal cavity above the supraglottic plan. In this session, subjects were not asked to produce buccal sounds because the larynx does not play a primary role in the production of clicks, beside adjustments (e.g. voicing, glottal closure, aspiration) relative to their phonemic function in a given language.

ii. Aerodynamics and Acoustics

Aerodynamic and acoustic signals were acquired with the EVA2 Workstation that allows simultaneous recording of acoustic signal, intraoral pressure, oral airflow and nasal airflow (Ghio and Teston, 2004). Acoustic waveform was obtained by EVA2 integrated microphone. Oral Airflow (Oaf), expressed in cubic decimeter per second (dm³/s), was collected using a flexible silicone mask pressed on the subjects' mouth. Nasal Airflow (Naf) also expressed in cubic decimeter per second (dm³/s), was obtained through a tube placed in the nostril. Intraoral pressure (Po), expressed in hectopascal (1 hPa = 1.02 cm H₂O), was obtained inserting a small tube into the pharynx through the nasal cavity, except for subject LJ where Po was obtained inserting the tube into the mouth; we took care to place the tube at the corner of the lip and perpendicular to the Airflow direction so Pressure and Oral airflow would not be mixed up. Thus, for this particular subject, pressure can only be analyzed for labial articulations, however the data shows unusual values up to 200hPa and more. We decided not to include nor analyze Po data of LJ. In some cases, beatboxers exhibited very high values for airflow and/or pressure that led to saturation.

C. ANALYSIS

i. Laryngoscopic Analysis

The analysis of video recordings will provide a description of laryngeal gestures and configurations. Based on the *Laryngeal Articulator Model* (Esling et al., 2019), we will focus on opening and closing gestures of Vocal Folds (VoF), Ventricular Folds (VeF), Aryepiglottic Folds (AeF) and Pharyngeal Walls (PhaW). We will also have a look at Tongue root, Epiglottis and Arytenoids movements. No data on buccal patterns was recorded because of the dissociation between buccal and pharyngeal and laryngeal cavities. Consequently, we assume VoF are in voicing position

since all beatboxers chose to add *humming*. It is similar to the voiced component of click accompaniments in languages where accompaniments are contrastive.

ii. Aerodynamic analysis

The analysis of aerodynamic signals will allow us to identify articulatory events. Indeed, since pressure depends on the volume, changes in intraoral pressure will provide informations about expansion/contraction in the vocal tract and consequently about the nature of beatboxing gestures (e.g. occlusion, trilling). Volume and pressure variations in the vocal tract set air in motion; more precisely they generate either egressive flows or ingressive flows at a given velocity. Description of Po, Oaf and Naf signals will shed light on oral, nasal and laryngeal gestures and their coordination in the production of beatboxing sounds and patterns.

iii. Acoustic analysis

The acoustic analysis will help us describe phonetic properties of beatboxed gestures and sounds since the waveform is synchronized with the aerodynamic recordings. First, we will give a description of the acoustic signal. Second, we propose to describe spectral characteristics of beatboxed sounds by means of spectrographic inspection.

3. RESULTS

A. KICK

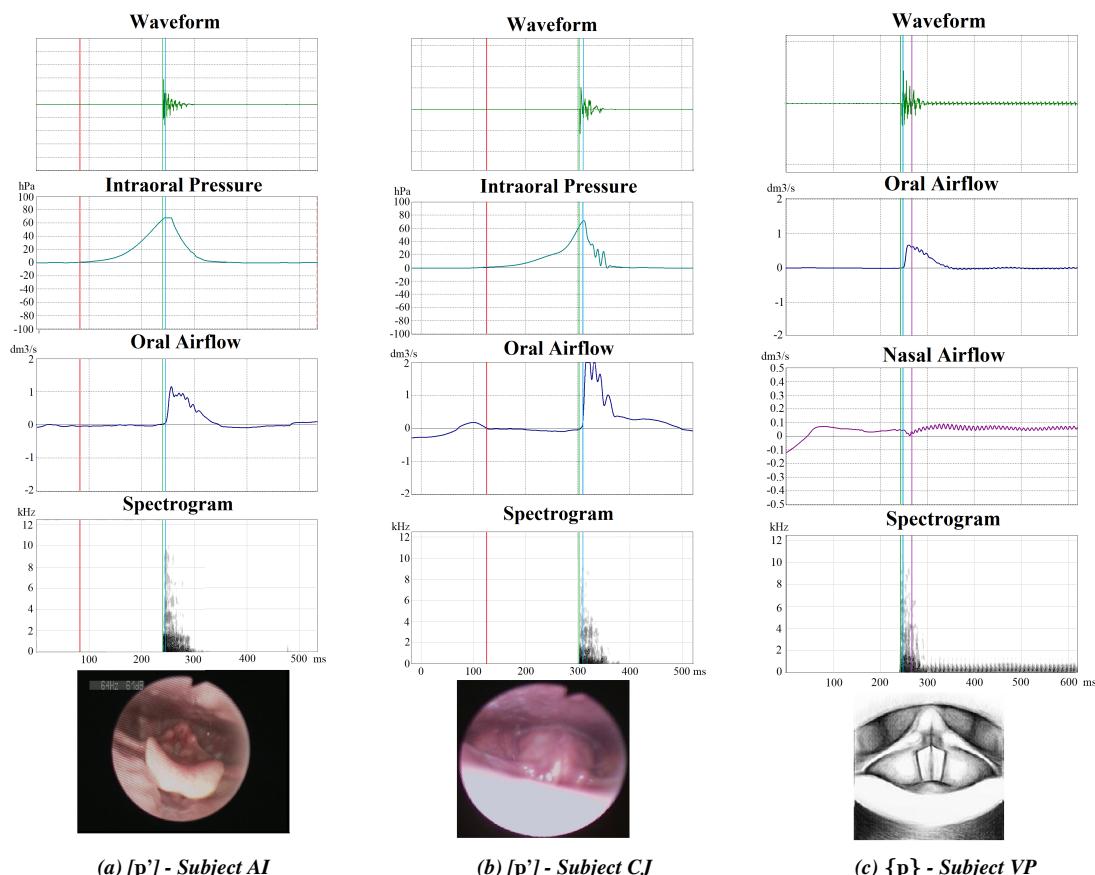


Figure 3: Production of Kick Drums (red line = occlusion onset, green line = transient onset, blue line = release offset). The fibroscopic images come from our data and are extracted at the burst onset, the drawing shows a voicing configuration to illustrate laryngeal configuration for humming and was taken from Esling et al., (2019) with their permission

The Classic Kick [p'] (Figure 3a) was produced as a voiceless bilabial stop initiated by a glottalic egressive airstream. Although individual differences were observed concerning the laryngeal configuration, the acoustic output displays striking similarities. We observed 3 cues confirming that the kick drum is a labial stop. First, Po is increasing while no Oaf nor any acoustic noise is observed on the signal, this is characteristic of a stop. Second, we noted a transient impulse on the waveform following the occlusion, which is typical of a burst; however, beatboxed bursts are atypical because no oral airflow is measured during the short time (i.e. approximately 5-10ms) of the burst. Because acoustic waves propagate quicker than air particles, it is not surprising to observe acoustic noise before Oaf increase. Third, the release shows an important concentration of energy around 0.5kHz or below. It is less than the range described by Halle et al. (1957) for labial stops. Furthermore, both acoustic and aerodynamic signals show an oscillation pattern typical of trilling. Indeed pressure and airflow peaks and valleys are in opposition. We attribute the trilling to the labial mucous membrane. Oral Airflow values are higher than in speech sound production, they range from 0.5dm³/s to 2dm³/s (i.e. saturation level of EVA2 workstation) and indicate that the airstream is egressive. It is clear that beatboxers use a glottalic airstream. We observed very high intraoral pressure values around 50hPa and up to 80hPa; it is surprising to observe such high values for bilabial ejectives given that pressure is correlated to the cavity volume behind the constriction point and the bigger the volume, the lower the pressure. It is possible that the vocal tract “soft” walls (e.g. cheeks, pharyngeal walls) are actively participating to air compression. High pressure is also explained by VoF closure and laryngeal raising. Indeed, fibroscopic images show adduction of the glottis, upward movement of the larynx and backward movements of the tongue root and epiglottis. Subject CJ produced an additional closure of the aryepiglottic tube.

The buccal kick {p} (Figure 3b) was produced as a bilabial stop initiated by a buccal egressive airstream, contrary to *clicks* that are ingressive. We found the same cues as [p'] confirming it is a labial stop. A silence followed by a low frequency burst, the spectrogram displays less energy below 0.5 kHz. We have no intraoral pressure for 4 subjects data on {p} because (1) pressure was acquired behind the dorsal constriction and (2) the velopharyngeal port is open to produce nasal voicing therefore Po should be low or null. Since for subject LJ pressure was acquired in the mouth we were able to measure pressure between the lips and the velic constriction; pressure values are close to 150 hPa. Again, we must carefully interpret the data since the tube was not orthogonal to the airflow. Airflow values are lower than for [p'], they range from 0.3dm³/s to 1dm³/s. Because subjects were asked to breathe or produce humming while beatboxing buccal sounds, Naf values show an egressive airstream and cues of voicing on the spectrogram of Figure 3b, where we chose to illustrate the larynx with a configuration representing voicing.

B. HI-HAT

The glottalic hi-hat [ts'] was produced as a coronal ejective affricate initiated by a glottalic egressive airstream. We identified various cues confirming the mode and place of articulation. The hi-hat begins with an occlusion (i.e. Po increase and acoustic silence). The waveform is characterized by an abrupt transient attack and followed by friction noise. Burst is short and atypical because airflow is null; Spectrographic analysis revealed the presence of energy between 0 and 12kHz. Spectrograms are identical between subjects: a spectral prominence around 1kHz and a less intense peak of energy around 5-8kHz are observed. High frequency peaks mean the anterior tube is short, it is typical of coronal stops and fricatives. The production of [ts'] is characterized by high pressure value ranging from 40hPa to 100hPa depending on the subject. Again, high pressure values indicate a glottalic mechanism. Similar to [p'], fibroscopic analysis revealed a glottal closure and laryngeal raising. We noted an additional closure of the AeF, for subject CJ. Oral Airflow is egressive values lie between 0.4 and 0.6 dm³/s; subject GA displayed values up to the saturation point (i.e. 2dm³/s).

The buccal {ts}, is a coronal affricate initiated by a buccal egressive airstream. Similar to [ts'], waveform exhibits an abrupt transient followed by friction noise. Spectrograms show the same acoustic characteristics but less energy compared to the glottalic hi-hat. Po was null, Oral Airflow values are low and range between 0.1 and 0.3 dm³/s. Egressive Naf was observed for humming.

C. VELAR SNARE

The non-buccal velar snare was produced as a voiceless pulmonic ingressive lateral velar affricate [$\downarrow k\acute{l}$] or as a segment composed of a voiceless glottalic ingressive velar stop and a voiceless velar pulmonic ingressive fricative [$\downarrow k\acute{l}$]. Whether it is a pulmonic or a glottalic airstream, the velar snare [$\downarrow k\acute{l}$] is produced by the opposite mechanism of an egressive affricate in speech: (1) tongue dorsum creates a complete closure in the velar zone, (2) the volume

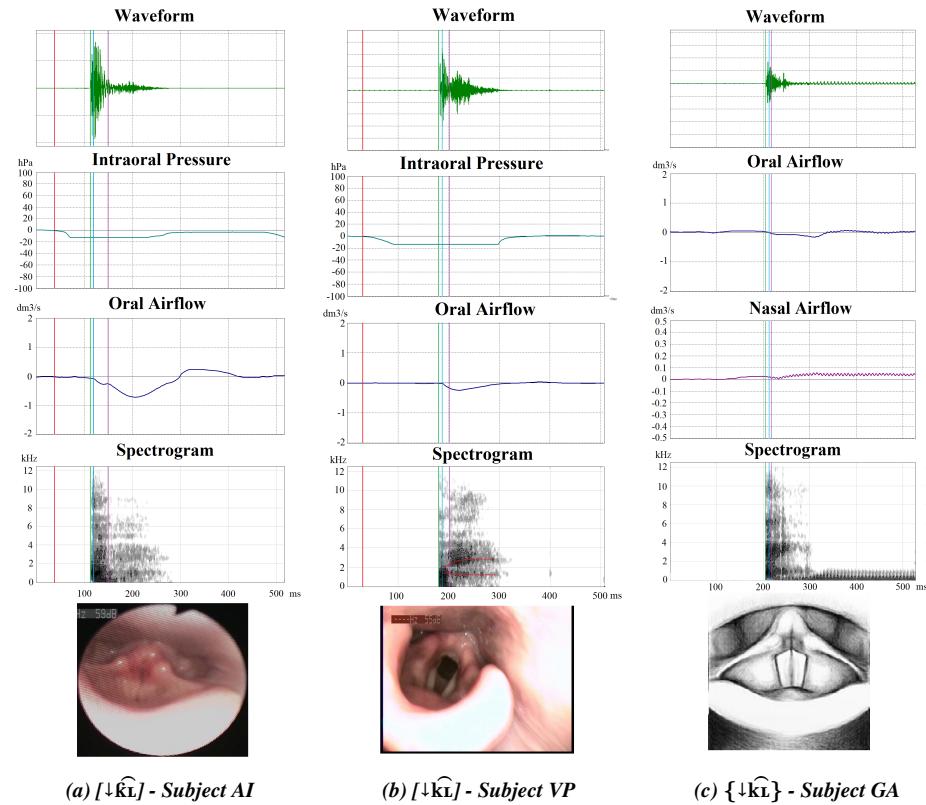


Figure 4: Production of ingressive velar snares (red line = occlusion onset, green line = transient onset, blue line = transient offset, purple line = frication onset). The fibroscopic images come from our data and are extracted at the burst onset, the drawing shows a voicing configuration and was taken from Esling et al., (2019) with their permission

behind the constriction point is expanding, leading pressure to decrease and (3) when the velar constriction is released, air is sucked in. Various gestures were identified, the first being the occlusion. Indeed, pressure decreases and Oaf is null. The burst is acoustically characterized by a transient attack followed by friction noise. No airflow was measured during the burst. Burst displays a peak of energy around 1kHz and resonances in the friction noise. Some tokens show approximation of the 2nd and 3rd resonance typical of velar stop formant transition. Both the pulmonic and glottalic affricate display negative values for intraoral pressure. Extreme pressure values led to difficulties when parameterizing pressure threshold. Pressure saturated at -12hPa for VP, AI, at -18hPa for LJ (only for labials) and at -100hPa for GA and CJ. Based on CJ and GA data, pressure reaches -30hPa to -50hPa and oral airflow ranges from -0.2 to $-2\text{dm}^3/\text{s}$. Fibroscopic images shows clearly that for speakers VP, LJ, CJ and GA, the glottis is opened during the entire gesture and the larynx is moving downward. Speaker VP's data show additional constriction of lateral pharyngeal wall and backward movement of tongue root giving a tubularized shape to the epilarynx. Speaker AI clearly uses a glottalic ingressive mechanism (i.e. implosive). The data shows (1) a closure of VoF (Vocal Folds) and VeF (Ventricular Folds), (2) laryngeal lowering and (3) glottal opening for the fricative.

We describe $\{\downarrow\widehat{k}\ell\}$ as a voiceless buccal ingressive lateral velar affricate. According to the pilot subject VP, the buccal snare is produced with 2 constriction points: a dental and a velar constriction. Contrary to *clicks* in speech, the dorsal constriction is released first. Based on the beatboxer's description, the air is trapped between constrictions while tongue dorsum is moving backward, increasing the volume between tongue tip and dorsum. Once the backward constriction is released air is sucked in. Aerodynamically, the data acquired is not sufficient to infer articulatory movements for the production of this sound. Indeed, Po is null, Oaf is null and Naf is egressive because of humming. Acoustically, we identified different phases on the waveform and spectrogram: a transient attack followed by friction noise. Subject VP showed additional noise of less amplitude on the signal and no energy above 5kHz before the attack transient.

D. LARYNGEAL SNARE

We call *Laryngeal Snare* sounds combining glottal and oral constrictions. The “Cough Snare” [ʔ̄h] is produced as pulmonic egressive epilaryngeal affricate. We identified 3 gestures: (1) (epi)glottal closure, (2) a voiced release and (3) aspiration noise.

Concerning the closing gesture, we observed differences between speakers on fibroscopic images. Subject VP (Figure 5a) uses a quasi-sphincteric mechanism of the epilaryngeal tube by means of pharyngeal narrowing, backward movement of the tongue root and the epiglottis hiding the glottal area. We assume that vocal folds are adducted. For AI (Figure 5b) we observed a full glottal closure of VoF and VeF while for GA (Figure 5c) and CJ we observed an aryepiglottic closure. Unfortunately we do not have data for subject LJ. The laryngeal constriction can either be described as a ventricular [?] or epiglottal [?] occlusion.

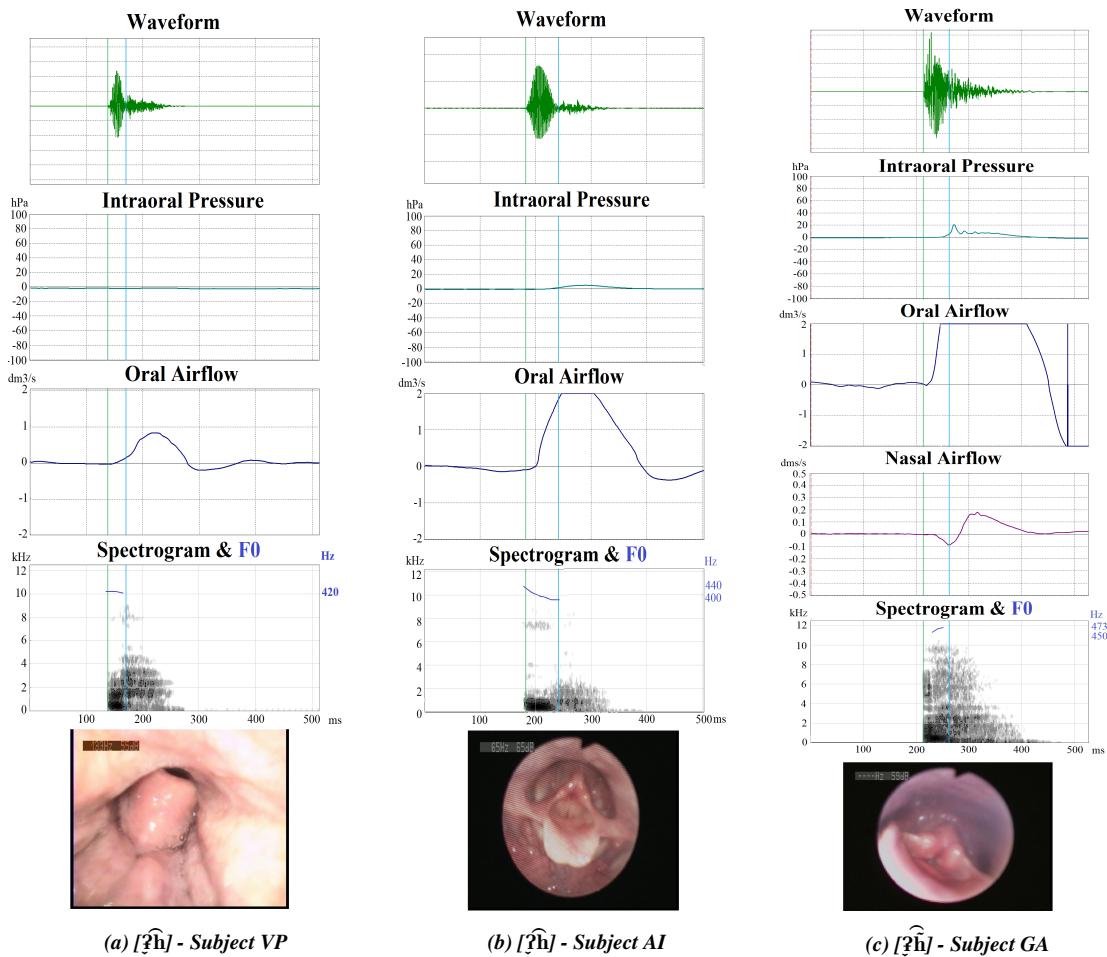


Figure 5: Production of the Cough Snare (green line = voiced release onset, blue line = aspiration onset). The fibroscopic image was extracted at the burst onset

A short period of voicing (approximately 20-30ms) is observed on the waveform and spectrogram. We interpret this as a voiced release of the epiglottal stop. Indeed, periodicity on the waveform and formants around 300-400Hz and 1200Hz are observed. It is clear that supralaryngeal cavities act like a filter similarly to vowel production in speech. F0 corresponds to F1, that is, 400Hz; it is comparable to pitch range of (breathy) falsetto voice (Esling et al., 2019), though this configuration (i.e. long and stretched VoF with no epilaryngeal constriction) is observed for subject AI only. For subject GA, we observe additional aryepiglottic constriction and a full aryepiglottic closure for CJ. Laryngeal configuration for both GA and CJ is more similar to whispery voice than falsetto voice. High F0 and epilaryngeal constriction can be surprising, though not incompatible. Indeed, isometric tension can produce high pitch values without stretching VoF. Isometric tension is a muscular contraction that does not change muscle length

when contraction is applied. Esling et al. (2019) gave a clear explanation of this mechanism applied to laryngeal musculature “Longitudinal tension and aryepiglottic tension [...] together produce an isometric tension whose net effect is to compress the airway while stretching the tight, thin vocal fold over a restricted length, generating a high frequency” (p.76). Though, an F0 of almost 500Hz cannot be explained only by stiff vocal folds. Given the high values of airflow (up to 2dm³/s), if subglottal pressure is sufficiently high along with the isometric tension, it is possible to reach 500Hz of fundamental frequency.

The Cough Snare ends with aspiration noise. We noted 3 cues confirming the presence of aspiration: (1) Po is null (i.e. absence of constriction), (2) resonances are observed on the spectrogram (i.e. formants) and (3) glottis is open wide. We observe a peak of Oaf of 1dm³/s for VP and saturation at 2dm³/s for subjects LJ, AI, CJ, GA. No pressure was measured except a 10hPa peak corresponding to the Oaf peak for subject AI; it may be due to a possible supralaryngeal constriction or narrowing that participates as a filter.

E. LIPS ROLL

The *Lips Roll* is a widespread beatboxing sound that is very difficult to acquire according to beatboxers. It is produced as an voiceless ingressive lateral bilabial trill that can be initiated either by a pulmonic or a buccal airstream. It is possible to start the pulmonic lips roll with a kick drum [$\downarrow B^1$] → [$\downarrow pB^1$]. For the purpose of the research, subjects were asked to initiate the lips roll with a kick drum. The pulmonic lips roll was similar for all speakers with subtle differences in the trilling. The kick drum, however, was produced either as pulmonic ingressive (Subject VP, LJ), a glottalic ingressive (Subject GA, AI - 6a) or a glottalic egressive (Subject CJ - Figure 6b). We identified 4 phases for the production of a lips roll initiated by a kick drum: (1) the occlusion, (2) the burst (transient), (3) a period of transition between the stop and the trilling gestures, (4) the trilling. For all subjects except CJ, intraoral pressure during the bilabial occlusion shows negative values around $-12hPa$ and down to $-95hPa$ (!). Airflow values range from $-0.2dm^3/s$ to $-2dm^3/s$. Concerning laryngoscopic images, subjects VP and LJ show an open glottis throughout the entire production, no signs of epilaryngeal constrictions were observed and both subjects exhibited a downward movement of the larynx. Subject VP showed slight pharyngeal narrowing. Subjects AI and GA used a glottalic ingressive mechanism to initiate the airflow. Figure 6a shows at the bottom the laryngeal configuration at burst onset, it shows the larynx is in a downward position and with a complete glottal closure and incomplete ventricular closure. Finally, subject CJ produced an ejective as shown on Figure 6b: pressure is up to 80hPa, Oaf to 1.4dm³/s. There are cues of epilaryngeal constriction and laryngeal raising on fibroscopic images. Spectral cues on the spectrogram are the same as kick drum (see 3A).

Let's turn now to the trilling mechanism. Whether initiated by a pulmonic or a buccal mechanism one, the trill is ingressive. Airflow is ranging from -0.1 to $-2dm^3/s$ (saturation level) for the pulmonic trill. The frequency of labial vibration is lower for subject CJ (37Hz) and subject VP (40Hz) and higher for subjects GA (45Hz), AI (48Hz) and LJ (50Hz). For VP and CJ we noticed periodicity on the wave form similar to what can be observed on Figure 6c (the two purple lines indicate the periodicity in the trilling). We do not attribute the periodic oscillation to voicing because either the glottis is wide open for the pulmonic Lips Roll or it is producing egressive nasal voicing for the buccal Lips Roll. We attribute the oscillation to the labial membrane vibration during the closed phase of lips. Indeed, similar to VoF vibration, the upper and lower membrane may vibrate under *Bernoulli's effect* during the closed phase of labial vibration. The period T ranges from 0.7-0.9 ms and creates a resonance F ($=1/T$) at 1000 or 1200Hz as illustrated on the spectrogram of Figure 6b.

Figure 6c illustrates the buccal lips roll. It is characterized by a general trilling of the lips and also by membrane vibration. Contrary to pulmonic lips roll, trilling frequency for both lips and membrane diminishes over time. Concerning lips vibration, the period T ranges from 8ms to 31ms depending on the subject. The frequency $F = 1/T$ ranges from 125 Hz to 32Hz. Concerning the membrane vibration (i.e. periodic oscillations Figure 6c) T is divided by 2 over time, except for subject LJ for whom T is constant. For the other subjects T ranges from 0.9 ms to 2.3ms and F ranges from 1111Hz to 434Hz. The vibration of the membrane can be seen on each spectrogram; moreover the diminution of the frequency can also be observed on the spectrogram. No pressure was measured, Oaf is on average 0.3dm³/s and 0.005dm³/s for Naf.

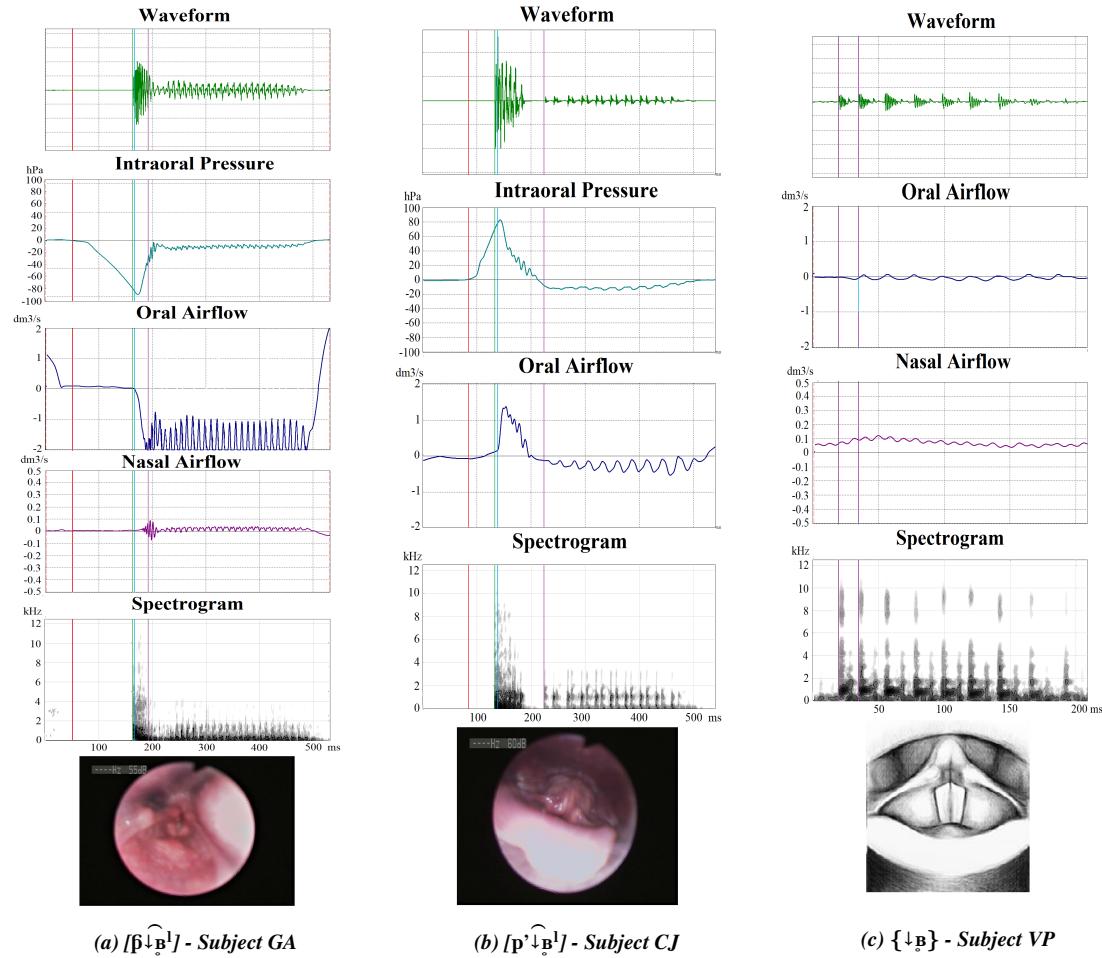


Figure 6: Production of Lips Rolls (red line = occlusion onset, green line = transient onset, blue line = transient offset, purple line = trilling onset). The fibroscopic images come from our data and are extracted at the burst onset, the drawing shows a voicing configuration and was taken from Esling et al., (2019) with their permission

4. DISCUSSION

A. AEROMECHANICS OF BEATBOXING

It is clear that beatboxers have an extended control on Vocal Tract capacities. For instance the production of the lips roll is remarkable. Trilling is a rather complex articulatory mechanism based on specific aerodynamic conditions. In order to initiate the trilling, articulators approximate to close the airway causing the pressure to increase behind the constriction. Air forces its way out creating an opening, pressure falls under the Bernoulli effect and articulators come back in contact closing the airway again. In the world's languages, trills are pulmonic and egressive, while in HBB, trills can be ingressive and either pulmonic or buccal (Blaylock et al., 2017). Beatboxers are able to manipulate aerodynamic conditions to generate an ingressive flow and initiate the trilling.

In the introduction we reminded the range of pressure values during speech production (0-15hPa) and the maximal and minimal range of possible pressure range in the vocal tract (-100hPa and +160hPa). Ranges of pressure for the production of beatboxing sounds are higher than speech production and imply greater volumic changes. *Boyle's Law* stipulates the inverse relationship between volume and pressure. Following the equations given in Catford (1977), we are able to evaluate volumic reduction in the vocal tract for the production of [ts'] where P_0 is 100hPa :

$$V_2 = \frac{P_1 \times V_1}{P_2} = \frac{1013 \times 140}{1113} = 127 \text{ (cm}^3\text{)}$$

V₂ is the reduced volume we want to calculate, P₁ is the initial pressure in an open vocal tract which is equal to the

atmospheric pressure (1013hPa) and V1 is a reference value given by Catford of the initial volume between the lips and the glottis. The volumic reduction is $140 - 127 = 13(cm^3)$.

Another interesting issue is the atypical bursts observed systematically in the data. Our analysis showed for stops and affricates that Oaf started after the acoustic transient attack. At first we thought it was a synchronization problem between signals during the pilot experiment, but all subjects exhibited the same characteristic and the duration of the observed gap was variable across sounds and subjects. Then, we interpreted the gap as the difference between sound wave velocity and particle velocity. Acoustic waves propagate at 343 m/s (34300cm/s) in air. The calculation of particle velocity is a little more difficult because it depends on several factors as pressure, density or gravitational acceleration. For speech purposes, Catford offers a simplified equation to evaluate particle velocity where u is particle velocity (cm/s) and p the pressure (in mmH_2O) before the constriction. Particle velocity for [ts'] with a pressure of 1019 mmH_2O (100hPa) :

$$u = 412\sqrt{p} = 412\sqrt{1030} = 412 \times 32 = 13184 \text{ (cm/s)}$$

Acoustic wave propagates at higher velocity than particles $34300 > 13184 \text{ cm/s}$. To our knowledge this gap between velocities was not observed in speech, thus we do not know if our interpretation is also observable in speech data.

B. BREATHING AND BEATBOXING

The Vocal Tract may be divided in four subsystems: pulmonic, laryngeal, velopharyngeal and oro-pharyngeal apparatus (Hixon et al., 2018). During speech, breathing is reorganized. While for quiet breathing the inhalation period is almost the same as exhalation, for speech production inhalation is shorter and exhalation is longer. The egressive airflow coming out of the lungs is then shaped into sounds by the other subsystems. The data presented here do not match speech breathing given the diversity of production mechanisms. Beatboxing is not produced during exhalation, in fact, pulmonic airflow is a production mechanism among others. Even though the data corresponds to sounds produced in isolation and not in beatboxing patterns, it raises the question of the reorganization of the breathing pattern. Beat Patterns (i.e. larger rhythmical beatboxed structures) are produced with an alternation between ingressive and egressive airflows. This alternation may be a strategy to have sufficient air in the lungs for gas exchange. It also raises the question about how beatboxers coordinate subsystems in order to insure their respiratory control.

Sounds were produced in isolation for the purpose of the study. However, beatboxers combine sounds together to produce larger rhythmical structures. The question about gestural coordination arises. Let's take as an example the following Beatboxed Rythmical Structure (i.e. *Beat Pattern*) [$\downarrow pb^1 \hat{ts'} \hat{?h} ts'$]. For subject GA we measured a range of -100 and $+100hPa$ to produce the lips roll and the closed hi-hat. In order to coordinate these 2 sounds, beatboxers would have to produce the specific Vocal Tract configuration to sufficiently lower the pressure and then change the configuration to sufficiently build pressure at a given beat (i.e. tempo). We think pressure values and ranges may differ from isolation when beatboxers are performing on stage. Similar to speech it is possible that artists overshoot or undershoot when they coordinate sounds. Indeed, for insuring a continuous production that complies with a specific rhythmical structure (i.e. metric, beat and timbre), beatboxers might undershoot articulatory targets or increase overlap between gestures. Our study on temporal reduction (Dehais-Underdown et al., 2020) showed that 6% of tokens ($n=1566$) displayed overlap cues, for example [ts] → [tsf] when neighboring labials such as kicks or snare drums. Overlapping occurs on few tokens in the acoustic data (*ibid*), however further MRI and aerodynamic investigations on Beat Patterns should provide extended knowledge on respiration, coordination and overlap.

5. CONCLUSION

Human Beatboxing is a great paradigm to explore the vocal tract capacities during production. Beatboxing artists have indeed an extended knowledge about their own vocal tract in order to combine different initiation and articulation mechanisms. HBB might bring new perspective on articulatory complexity and phonetic diversity (i.e. "universals") in the world's languages.

Another interesting fact is the absence of vocal lesions or injuries; given that beatboxers use more powerful mechanisms, we expect them to have a strategy to stay healthy. The understanding of such mechanisms might bring a useful and an original contribution for speech pathology.

6. ACKNOWLEDGMENT

We thank *Hôpital Foch* for their collaboration, especially concerning the help to obtain RCB-ID (n° 2020-A00246-33) and also the Labex EFL (Empirical foundations in linguistics), Axe 1 phonetic and phonological complexity.

REFERENCES

- Association, I. P. (1999). *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet*. Cambridge University Press.
- Ball, M. J., Howard, S. J., & Miller, K. (2018). Revisions to the extIPA chart. *Journal of the International Phonetic Association*, 48(2), 155–164. <https://doi.org/10.1017/S0025100317000147>
- Bernhardt, B., & Ball, M. J. (1993). Characteristics of Atypical Speech currently not included in the Extensions to the IPA. *Journal of the International Phonetic Association*, 23(1), 35–38. <https://doi.org/10.1017/S0025100300004771>
- Blaylock, R., Patil, N., Greer, T., & Narayanan, S. S. (2017). Sounds of the human vocal tract. *INTERSPEECH*, 2287–2291.
- Catford, J. C. et al. (1977). *Fundamental problems in phonetics*. Indiana University Press.
- De Torcy, T., Clouet, A., Pillot-Loiseau, C., Vaissiere, J., Brasnu, D., & Crevier-Buchman, L. (2014). A video-fiberscopic study of laryngopharyngeal behaviour in the human beatbox. *Logopedics Phoniatrics Vocology*, 39(1), 38–48.
- Dehais-Underdown, A., Crevier-Buchman, L., & Demolin, D. (2019). Acoustico-physiological coordination in the human beatbox: A pilot study on the beatboxed classic kick drum. *19th International Congress of Phonetic Sciences*.
- Dehais-Underdown, A., Vignes, P., Crevier-Buchman, L., & Demolin, D. (2020). Human beatboxing: A preliminary study on temporal reduction. *12th International Seminar on Speech Production*.
- Eklund, R. (2008). Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *Journal of the International Phonetic Association*, 38(3), 235–324.
- Esling, J., Moisik, S., Benner, A., & Crevier-Buchman, L. (2019). *Voice quality : The laryngeal articulator model*. Cambridge University Press. <https://doi.org/10.1017/9781108696555>
- Ghio, A., & Teston, B. (2004). Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies. *International Conference on Voice Physiology and Biomechanics*, 55–58.
- Gick, B., Wilson, I., & Derrick, D. (2012). *Articulatory phonetics*. John Wiley & Sons.
- Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of phonetics*, 29(4), 383–406.
- Hixon, T. J., Weismer, G., & Hoit, J. D. (2018). *Preclinical speech science: Anatomy, physiology, acoustics, and perception*. Plural Publishing.
- Paroni, A., Henrich Bernardoni, N., Savariaux, C., Loevenbruck, H., Calabrese, P., Pellegrini, T., Mouysset, S., & Gerber, S. (2021). Vocal drum sounds in human beatboxing: An acoustic and articulatory exploration using electromagnetic articulography. *The Journal of the Acoustical Society of America*, 149(1), 191–206.
- Proctor, M., Bresch, E., Byrd, D., Nayak, K., & Narayanan, S. (2013). Paralinguistic mechanisms of production in human “beatboxing”: A real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 133(2), 1043–1054.
- Sapthavee, A., Yi, P., & Sims, H. S. (2014). Functional endoscopic analysis of beatbox performers. *Journal of Voice*, 28(3), 328–331.