Human Beatboxing : pushing the boundaries of human voice production

Annalisa Paroni

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement
Spécialité : CIA - Ingénierie de la Cognition, de l'interaction, de l'Apprentissage et de la création
Unité de recherche : Grenoble Images Parole Signal Automatique

Le Human Beatbox : aux limites des frontières de la production vocale humaine

Human Beatboxing: pushing the boundaries of human voice production

Présentée par :

Annalisa PARONI

Direction de thèse :

Nathalie HENRICH BERNARDONI Directrice de Recherche, CNRS	Directrice de thèse
Hélène LOEVENBRUCK Chercheuse, Université Grenoble Alpes	Co-directrice de thèse

Rapporteurs :

CLAUDIO ZMARICH
Chargé de recherche, ISTC
STEN TERNSTROM
Professeur, KTH Royal Institute of Technology

Thèse soutenue publiquement le **5 septembre 2022**, devant le jury composé de :

NATHALIE HENRICH BERNARDONI Directeur de recherche, CNRS DELEGATION ALPES	Directrice de thèse
HELENE LOEVENBRUCK Directeur de recherche, CNRS DELEGATION ALPES	Co-directrice de thèse
SUSANNE FUCHS Directeur de recherche, Leibniz - ZAS	Examinatrice
CLAIRE PILLOT-LOISEAU Maître de conférences HDR, UNIVERSITE PARIS 3 - SORBONNE NOUVELLE	Examinatrice
SHRIKANTH NARAYANAN Professeur, University of California, Los Angeles	Examinateur
BENJAMIN LECOUTEUX Maître de conférences HDR, UNIVERSITE GRENOBLE ALPES	Examinateur (Président du Jury)
CLAUDIO ZMARICH Chargé de recherche, ISTC	Rapporteur
STEN TERNSTROM Professeur, KTH Royal Institute of Technology	Rapporteur

Invités :

MARIAPAOLA D'IMPERIO
Professeur émérite, Rutgers, The State University of New Jersey

Considerate la vostra semenza:
fatti non foste a viver come bruti
ma per seguire virtute e
caunoscenza

Dante Alighieri
Divina Commedia, Inferno,
canto XXVI

A vous, mes sœurs...

Acknowledgements

This work would not have seen the light of day had it not been for the perseverance and the continuous support of my supervisors: you truly have all my gratitude. Merci, Nathalie, for your guidance, your patience, your wisdom, your scientific knowledge, and the friendship outside and at work. I could not have dreamed of a better supervisor. Merci, Hélène, for your constant encouragement, for believing in me and in my abilities, for always putting things into perspective. Thank you to all the members of the committee who have accepted to examine and review this manuscript, some on a more short notice than others. First of all, grazie, Claudio, for setting all of this in motion so many years ago: I hope that the disappointment I caused you in choosing beatboxing over classical singing as a field of research has faded. Merci, Claire, for having followed quite closely this journey of mine all along and for your precious suggestions and observations. Danke, Susanne, for supporting our work and defending it against reviewer n.2! Thank you, Shri, for opening us the doors of your lab. I hope that some time in the future we will be able to accomplish this collaboration. Tack, Sten: I hope that you will find that beatboxing is at least as interesting as vocal imitations. Grazie, Mariapaola, and merci, Benjamin.

Many thanks to the colleagues that have provided technical and scientific support during the years: Pierre Baraduc, Christophe Savariaux, Julien Frère, Silvain Gerber, Coriandre Vilain, Thomas Pellegrini, Sandrine Mouysset, Pascale Calabrese, Maëva Garnier, Thomas Hueber, Giovanni De Pau, Anne Vilain.

Thank you to all the colleagues at GIPSA-lab, that has been my lab for way longer than the duration of a Ph.D. program! In particular, a huge thank you to Anneke Slis and Nadège Rochat for sharing long and busy days at the lab and for all the chats, the videos, and the fake articles shared over the years. I miss you!

A special thanks to Reed Blaylock and Alexis Dehais Underdown for the insightful discussions on beatboxing and for sharing with me this unique experience of building our research skills on such a bizarre, yet fascinating topic that is beatboxing.

A very very special thank you to Timothée Maison for our discussions on the wonders of the voice that only we can understand. Thank you to Clémentine Darj, because it is never enough. Thank you to Rémi Blandin, because, without you, I would never ever have gotten over moving from the seaside to the mountains.

Last but not least, thank you to my family, for supporting me in the many changes of path I had over the years, and thank you to all of my friends here in Grenoble, Lyon, and Montpellier that have kept me in touch with reality and myself. In particular, thank you, my sisters. It has been an extraordinary grace to walk together.

Contents

List of Figures

List of Tables

List of Abbreviations

GIPSA-lab	Laboratoire Grenoble Image Parole Signal Automatique
VF	Vocal Folds
VT	Vocal Tract
HBB	Human Beatboxing
P	Kick
t	Hi-hat
K	Rimshot
EMA	Electromagnetic Articulography
RIP	Respiratory Inductive Plethysmography
sEMG	surface Electromyography
EGG	Electroglottography
C1	Corpus number 1
C2	Corpus number 2
C3	Corpus number 3
PS	Pilot subject
S01	Subject number 1
S02	Subject number 2
S03	Subject number 3
S04	Subject number 4
S05	Subject number 5

Introduction

The voice... What a marvellous gift we are born with! The vocal organs... What an incredible instrument our body comes with!

From the moment we leave the womb, our voice signals our presence to the world and expresses our displeasure for being forced out of the comfort and warmth of mummy's tummy. Soon enough we learn that this voice is interesting and fun. We take pleasure experimenting with it: "babababa... mamamama...". And mum says: "Oh, yes honey! Papa! Mama!". All of a sudden, we realize that some of our vocal play has no particular meaning and some has a meaning and we develop our ability to speak. We learn that we can use our voice to interact with other people, or to express our sadness and our joy, we can tell stories, we can laugh, we can call our little furry friend, we can fake the sound of an engine while playing with our toy car, we can do a lot of things with it. Oh, but wait: we also use our voice to sing and act, and we don't use it the same way as we do when we speak... And then we are bored with our day and we go on YouTube and stumble across a video of this guy who uses his voice to make music, but he is not singing. He makes all these bizarre noises, but, hey, it sounds amazing! How does he do that? Does he have a microphone in his throat or what? How does he produce multiple sounds simultaneously? Does he have a whole band in his throat? Does he even breathe? Is he even human? Yes, this is beatboxing... human beatboxing.

Almost a decade ago, when I was undergoing my training as a speech therapist and was interested in more conventional singing styles (western operatic singing, go figure...), I had the chance of being introduced to this exceptional vocal art that is human beatboxing by the person who would later become my mentor and my research director. I was completely smitten and fascinated with this peculiar use of the voice. However, when I started to look into the literature, I was quickly frustrated by the inexplicable lack of scientific investigations on beatboxing and the missed opportunity for the Scientific Community to comprehend unusual and obviously very skilful sound production mechanisms that can very well be at the limits of the physiological capabilities of the human voice organs. This dearth of studies is certainly not due to technical issues: we do have advanced technologies that support the exploration of voice production and are widely exploited for the study of other kinds of voice production. Suffice it to think of the wealth of publications on speech and classical singing, for instance. This had to change. At GIPSA-lab, Nathalie Henrich Bernardoni was willing to start a long-term multidisciplinary project 'Beatbox' completely dedicated to the scientific study of this vocal art. Being one of the most prominent laboratories in the world that investigates speech, GIPSA-lab has all the equipment and expertise necessary to sustain this project. Further, the French beatboxing community is very active and exceptionally high-level, with multiple world champions. Beatboxing is

becoming more and more familiar to the general public, especially the younger population. In the springtime of this year 2022, a movie has seen the light of day in all the French movie theaters, starring a famous French beatboxer (MB14) as the protagonist. As for my part in the research on beatboxing at GIPSA-lab, I was involved in an ultrasound exploration of the lingual gestures of the repertoire of a professional beatboxer, that constituted the work for my bachelor dissertation in Speech and Language Therapy. We were able to observe some speech-like and some non speech-like articulatory mechanisms. Later, I was part of an acoustic study, presented in my dissertation of the first year of master program in Linguistics. We found that beatboxing sounds are longer and more intense than speech sounds, with higher intra-oral pressure and time derivative of the oral flow rate. I was then involved in two articulatory studies and one breathing study presented in the dissertations of my two master's degrees in Linguistics and Cognitive Sciences. Again, we observed speech-like and non speech-like articulatory mechanisms, as well as indications that breathing is used differently in speech and in beatboxing. However, these studies were very limited. Yet, they painted an interesting picture of unusual articulatory mechanisms and breathing behavior, different from those of speech, that paved the way for the more extensive work carried out during the years of my Ph.D. program, that I am going to present in this dissertation.

When I enrolled in the Ph.D. program, other than fundamental issues, I was still interested in the use of beatboxing for clinical applications, as a vocal play that emphasizes the use of the vocal instrument. The first results we obtained and in general the available literature pointed in the direction of a greater solicitation of the articulators (lips, tongue, and larynx) and the need for articulatory accuracy in beatboxing production. During various conversations with different beatboxers, I was told that they can feel that their oral muscles are particularly strong, some to the point that they think the volume of their tongue has grown over the years of practicing beatboxing. Weather this is true or not (yet again, no study has ever tackled the issue), from the available knowledge at the time it seemed reasonable to envision a set of exercises based on beatboxing to be used in speech therapy. In particular, my interest was set on Orofacial Myofunctional Disorders (OMDs). OMDs are associated with abnormal movement patterns of the facial and oral structures and imbalanced muscle strength. Usually, they are treated with a combination of maxillo-facial surgery and neuromuscular re-education, but more often than not neuromuscular re-education and orthodontics are enough. During my training as a speech therapist, I remember thinking that the myofunctional therapy programs I was being familiarized with sometimes asked for some very boring exercises to reinforce lingual and labial muscles. My thought was that beatboxing could provide a more interesting ground for muscle strengthening and articulatory precision. However, given the nature of this type of disorders, I felt the need to better understand the details of the production mechanisms of beatboxing before addressing the design of a speech therapy protocol that includes beatboxing.

Thus, the research project I would work on during my Ph.D. program aimed at investigating if the use of exercises derived from the practice of human beatboxing would facilitate the rehabilitation of articulatory precision or accelerate the dynamic muscular development of the orofacial sphere within the framework of a myofunctional therapy. In particular, we were interested in identifying a set of beatboxing sounds produced via articulatory and muscular mechanisms suitable for OMDs re-education exercises. In order to answer this question, the thesis work envisaged was structured in three research axes:

1. acoustic, phonetic, and articulatory description of the percussion sounds of human beatboxing compared to the occlusive and fricative consonants of speech;
2. development and evaluation of exercises adapted to the re-education of OMDs;
3. evaluation of the impact of these exercises in a clinical context.

Very quickly, we realized that the complexity of beatboxing sound production would require more time and effort to be investigated than what we had anticipated and that is available in the context of a Ph.D. program. Moreover, being more and more acquainted with the exceptional production mechanisms of beatboxing and more new questions being opened than old ones being answered, I progressively diverted my interest from clinical applications of beatboxing to the fundamental issues of phonetic comprehension of this vocal art.

Therefore, we conducted our investigations following two perspectives. On the one hand, we compared the production mechanisms of HBB sounds to those of similar speech sounds, in order to identify what is specific to HBB and what is similar to speech. On the other hand, we focused on the production mechanisms of HBB sounds.

Axis 1 Comparison between HBB and speech (articulatory dynamics, acoustics, ventilation, muscular activation).

Axis 2 Investigation of beatboxing-specific articulatory dynamics, acoustics, ventilation, muscular activation.

In particular, we focused on three main research questions (**Q**) for which we have formulated specific hypotheses (**H**):

Question 1 *Universality of sound production*: Are the articulatory strategies for producing the basic effects (kick, snare, hi-hat, rimshot, cymbal) and their acoustic results shared by beatboxers or are they beatboxer-specific?

Q1–H1 The acoustic results are similar according to the classes of beatbox sounds studied.

Q1–H2 For some basic effects such as "classic kick", "classic snare drum", "k-snare", "closed and open hi-hat", the articulatory strategy in terms of place of articulation and articulatory kinematics will be similar for all beatboxers.

Q1–H3 For less common effects or more complex techniques (e.g. "fast hi-hat"), each beatboxer develops their own articulatory strategy.

Question 2 *Comparison HBB-speech*: What are the similarities and differences between speech and beatboxing for a consonant and its beatboxed counterpart, in the case of the three consonants /p,t,k/ and the associated HBB sound categories kick/hi-hat/rimshot?

Q2–H1 The place of articulation is similar between speech and HBB.

Q2–H2 The articulatory kinematics for the lips and tongue differs between speech and HBB. For instance, a bilabial pulmonic occlusive in speech (/p/) becomes a bilabial ejective (kick).

Q2–H3 Facial muscle activation (measurable by surface EMG) is less in speech than in HBB.

Q2–H4 There is less right dominance of orofacial muscle activation in HBB compared to speech (for right-handed beatboxers).

Q2–H5 The breathing behavior is substantially differentiated between HBB and speech.

Question 3 *Voicing*: Does the addition of voicing (in the case of HBB: humming vs non-humming modification) change the HBB articulatory behavior in a different way than speech?

Q3–H1 Compared to a non-humming technique, the humming technique changes the articulatory behavior at the level of the flesh points studied (change in articulatory mechanism), reflecting the change in initiation mechanism. This is not the case for speech, when voicing is added (e.g. [p, b]).

One of the corpora used in the work presented in this dissertation, corpus C2 (section 3.2.2) was specifically designed to investigate these questions. However, the work presented tackles only a subset of the formulated hypotheses. In particular, **Q1–H3** and **Q2–H4** are left for future investigation.

Throughout this manuscript, we will outline the basics of human voice production for linguistic, para-linguistic and non-linguistic purposes (Chapter 1). We will then discuss

human beatboxing as a vocal art and as a field of voice research (Chapter 2). In Chapter 3, the technical aspects of this research will be presented. The results are organized as follows: Chapter 4 focuses on axis 2, characterizing the production mechanisms of 5 categories of drum sounds; Chapter 5 focuses on axis 1, investigating similarities and differences between corresponding beatboxing and speech sounds; Chapter 6 focuses on **Q3**, investigating a peculiar use of the vocal tract typical of beatboxing. Lastly, a general overview on the main findings of this work is presented, as well as future lines of investigation.

Part I

Theoretical background

Chapter 1

Human voice production

Contents

1.1 Anatomy and physiology

Voice production is a prodigy of the human body. Last to appear in the evolution of the vocal tract (VT), it shares anatomical structures with other primary functions, such as breathing, swallowing, and nutrition. Furthermore, voice production largely relies on breath supply.

1.1.1 Breathing

For the purpose of this work, we will focus on the first process, ventilation, i.e. the transport of air from the nose or mouth to the lungs and vice versa. Thereafter, we will use the terms respiration, breathing and breath to address this process. The transport of air is ensured by the two acts of breathing: inhalation or inspiration and exhalation or expiration, managed by the action of several muscles and the elastic properties of the structures involved in

breathing. Through inhalation, air is introduced into the respiratory system through the expansion of the volume inside the thoracic cavity. From an aerodynamic standpoint, this causes a decrease in pressure inside the lungs compared to the outside. The introduction of air rebalances this pressure difference (Titze & Martin, 1998). Exhalation, on the other hand, is a consequence of a decrease in thoracic volume, which causes an increase in pressure inside the lungs relative to the exterior environment and results in the expulsion of air outside of the respiratory system.

The respiratory system consists of all the ducts that allow air to enter and leave the lungs and the lungs themselves. It is conventionally divided into two parts (Fig. 1.1): the upper tract or upper airways, also called the vocal tract (Titze & Martin, 1998), that include the nasal cavities, the oral cavity, the pharynx, and the larynx; and the lower tract or lower airways, that include the trachea, the bronchi, the bronchioles, and the alveoli.

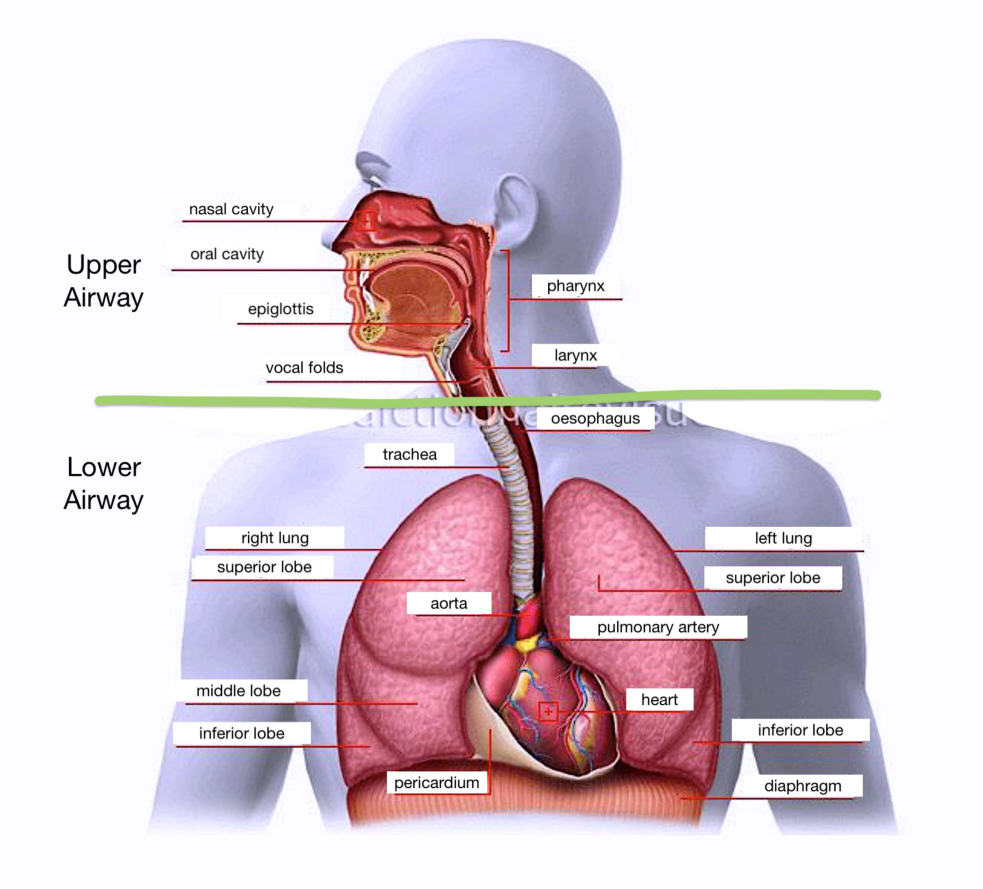


Figure 1.1: The respiratory system.

The structures of the upper airways are located between the cranial base, cervical spine, and sternoclavicular plane, and are separated from the lower airways by the larynx. The structures of the lower airways are located within the rib cage and are separated from the

abdominal cavity by the diaphragm. As their placement suggests, the upper airways are not dedicated solely to respiratory function, but are used for the execution of multiple very different functions, such as swallowing, phonation and verbal articulation. The lower airways on the other hand are dedicated solely to breathing.

The air travels through the airways by means of pressure differences between the inside and outside of the body, generated by volume variations of the thoracic cavity. In turn, these are generally obtained by the action of different respiratory muscles, depending on the inhalation and the exhalation phase, but also according to the amplitude of thoracic volume variation required.

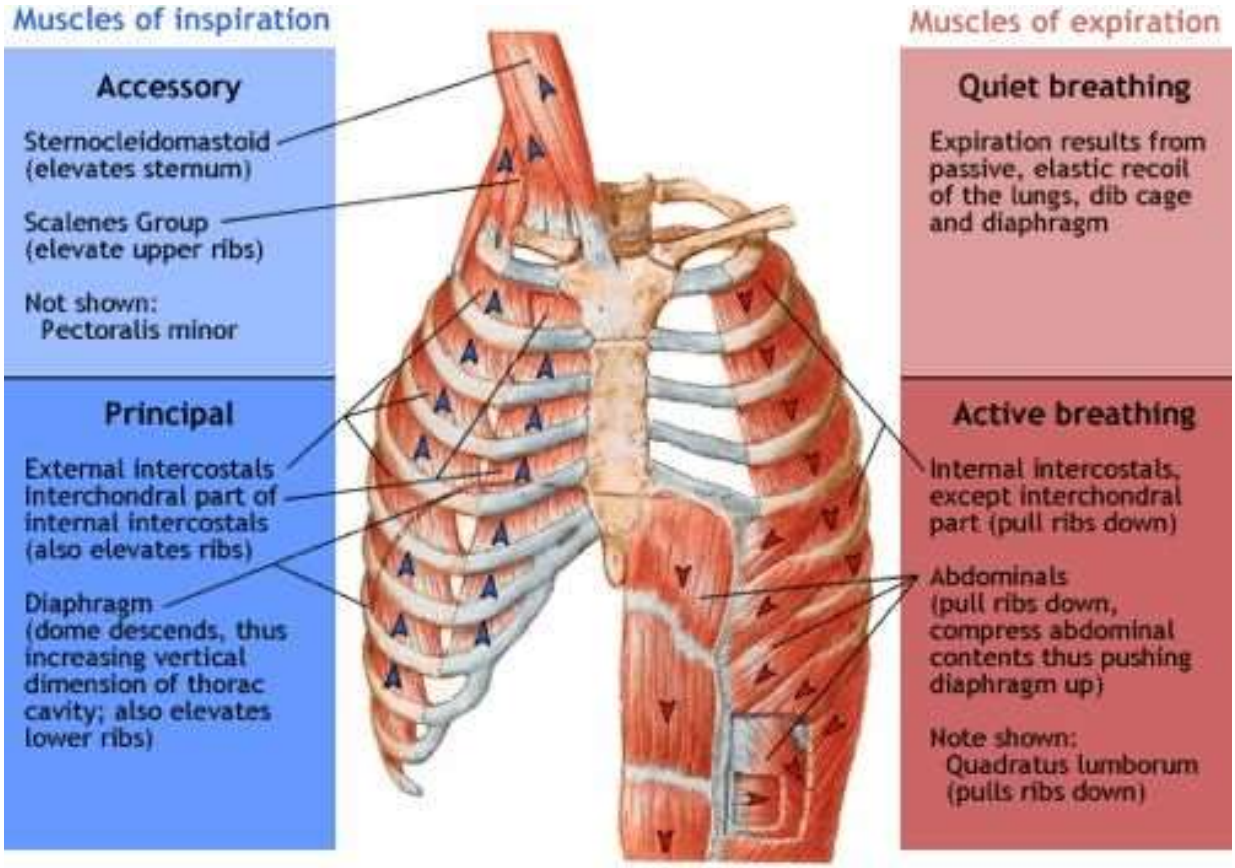


Figure 1.2: The respiratory muscles.

In a resting breathing condition, the thoracic expansion that takes place during inspiration is ensured by the action of the diaphragm in particular, but also of the parasternal or external intercostal muscles (Fig. 1.2): the diaphragm contracts and flattens, resulting in an increase in the space available at the base of the thoracic cavity; the external intercostal muscles contract and slightly raise the ribs, causing an increase in the cross-sectional area of the thoracic cage. The decrease in thoracic volume, on the other hand, is a passive pro-

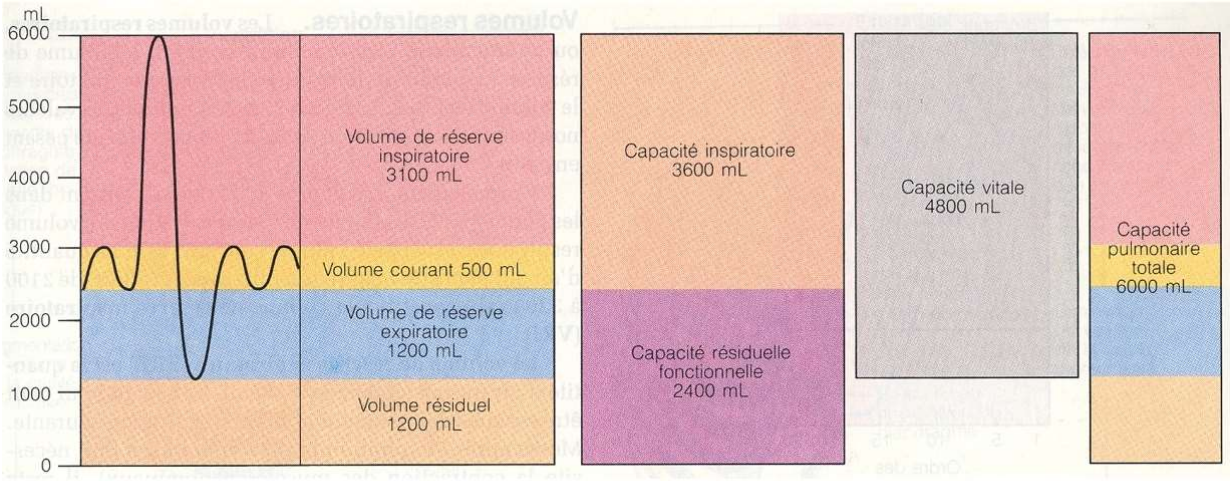
cess that occurs during the exhalation phase: the inspiratory muscles cease their activity and the elastic recoil forces of the muscles and lungs return these structures to their resting position and the volume of the thoracic cavity to its initial volume.

The purpose of respiration is to ensure the supply of O_2 and the removal of CO_2 . However, the quantities of these exchanged substances vary according to the demands of the organism and the variations in respiratory volume at rest may not be enough to meet the needs of the organism. In this case, another type of respiration comes into play, i.e. forced breathing. Larger and faster volume variations are ensured by the action of several muscles, both during the inhalation and exhalation phases. The main active muscles in the forced inspiration phase are the same as for inspiration at rest, i.e. the diaphragm and the external intercostals. Their action is reinforced by the intervention of other muscles, called accessory muscles, such as the scalenes and the sternocleidomastoid. The contraction of these muscles increases the volume of the upper part of the thoracic cavity. As the recoil forces are no longer sufficient, during forced respiration the exhalation phase also becomes active. Several muscles contract in order to compress the thoracic cavity and expel air out of the airways. The internal intercostal muscles reduce the volume in the upper part of the thoracic cavity by lowering the ribs, the abdominal muscles (rectus abdominis, internal and external oblique, transverse) contract to increase the rise of the diaphragm and the volume is reduced in the lower part of the thoracic cavity. In general, during inspiration, there is an increase in thoracic volume, first in the lower part of the thoracic cavity, by means of the activity of the diaphragm, and then with a more forced respiration, progressively higher. The expansion of the thorax is not uniform. The expansion and contraction movements of the thoracic cage are the result of the interaction between the morphology of the ribs, the costo-vertebral joints and the respiratory muscles. This means that during inspiration, the expansion of the upper part of the thorax is mainly in the anterior-posterior direction, while in the lower part of the thorax the expansion is more medial-lateral (Bastir et al., 2017). Therefore, a distinction is made between a pulmonary compartment and a diaphragmatic compartment of the thorax.

Vital capacity of the lungs is defined as the largest volume of air that can be exchanged within the lungs (Huber & Stathopoulos, 2015) during a forced inhalation and then a forced exhalation. This constitute only a portion of the total lung volume (Titze & Martin, 1998). Formally, vital capacity is defined as the amount of air that can be exhaled following a maximal inspiration (Huber & Stathopoulos, 2015). Forced vital capacity is similar to vital capacity and is defined by the amount of air that can be exhaled as forcefully as possible following a maximal inspiration (Huber & Stathopoulos, 2015).

Vital capacity is divided into three components (Titze & Martin, 1998): expiratory reserve volume, tidal volume and inspiratory reserve volume. Tidal volume is defined as the amount of air inspired and exhaled during breathing at rest. It represents only 10-15% of vital capacity during breathing at rest, but increases with physical activity (increased

demand of O_2 and production of CO_2) until it reaches 50% of vital capacity during intense activity (Dickson & Maue-Dickson, 1982). The expiratory reserve volume represents the volume of air that is exhaled between the end of a quiet exhalation and the end of a forced exhalation. The inspiratory reserve volume is the volume of air that is inspired between the end of a quiet inspiration and the end of a forced inspiration. The inspiratory capacity is the maximum amount of air that can be inspired after a normal expiration, i.e. vital capacity plus inspiratory reserve volume.



Anatomie et Physiologie Humaines, Elaine N. Marieb; De Boeck Université.

Figure 1.3: Pulmonary air volumes and capacities.

Residual volume is the volume of air remaining in the lungs after a forced expiration (Huber & Stathopoulos, 2015). This is a non-mobilizable amount of air. Functional residual capacity represents the volume of air in the lungs at the end of a quiet exhalation. Total lung capacity is the maximum amount of air contained in the lungs. A respiratory cycle consists of an inspiration and an expiration and is characterized by its total duration, equal to the sum of the inspiratory and expiratory durations, its amplitude, and the tidal volume. The inspiratory and expiratory durations are not perfectly equal, but are very similar and the respiratory cycle shows a certain symmetry. (Fig. 1.4) The ventilatory frequency expresses the number of ventilatory cycles that take place in one minute.

These quantities vary in the same individual according to multiple factors, including for instance emotional state and physical activity. Further, they vary greatly from one individual to another. The respiratory frequency can vary from 6 to 31 cycles per minute; the resting tidal volumes can vary between 442 ml and 1549 ml (Dejours et al., 1961); the inspiratory and expiratory time durations vary, although inhalation is always shorter than exhalation for a single respiratory cycle. An infinite number of combinations of tidal volume and respiratory frequency can result in the same flow rate. Each individual has

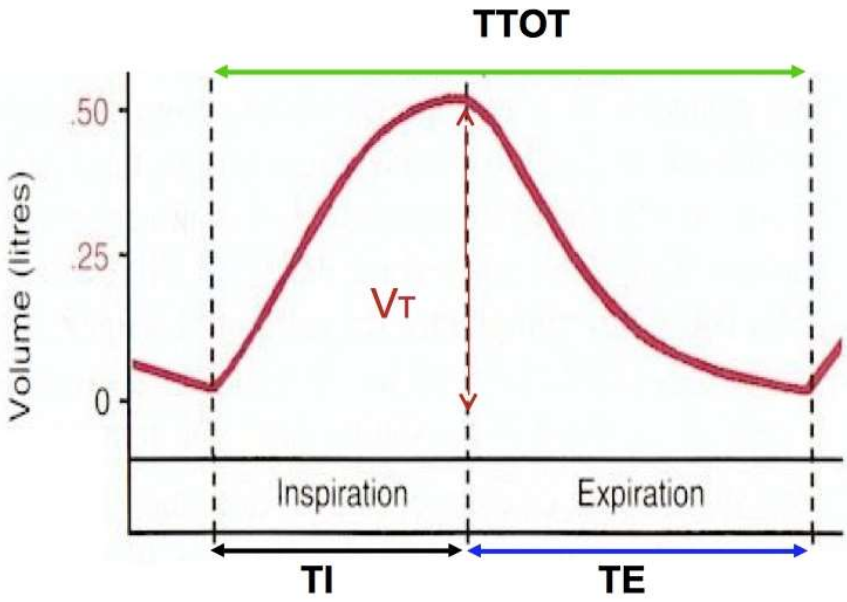


Figure 1.4: Respiratory cycle at rest. Figure courtesy of Pascale Calabrese.

their own combination of these two quantities and a form of respiratory movements which is peculiar to them. In essence, each individual has their own respiratory mode or *respiratory personality* ("personnalité ventilatoire", Dejours et al., 1961). Among the infinite number of combinations of tidal volume, respiratory frequency, and airflow pattern, each individual chooses a particular pattern (Shea & Guz, 1992) which is prone to be maintained in different conditions, including hypoxia (Eisele et al., 1992). Multiple studies have shown that each individual has a unique ventilatory mode. The shape of the airflow is unique to each person, depending on the intrinsic properties of their respiratory system (Benchetrit et al., 1989; Besleaga et al., 2016) and would be reproducible over time (at a distance of 4-5 years), despite changes in habits such as smoking or variations in body weight (Benchetrit et al., 1989).

1.1.2 Phonation and the Source

As previously mentioned, breathing and voice production are intertwined. While for respiration, opening and closing movements of the vocal folds may suffice, swallowing, voice production in general and more so speech require much more complex laryngeal movements. The vocal folds have an additional function to that of a protective valve of the lower airways. They are the *organ of phonation*. The larynx plays a fundamental role in regulating the airflow coming from the lungs (Leanderson & Sundberg, 1988). According to the Source-Filter theory (Fant, 1970), when the vocal folds are adducted, the column of air exiting the lungs creates an increase in subglottal pressure, until it forces the vocal

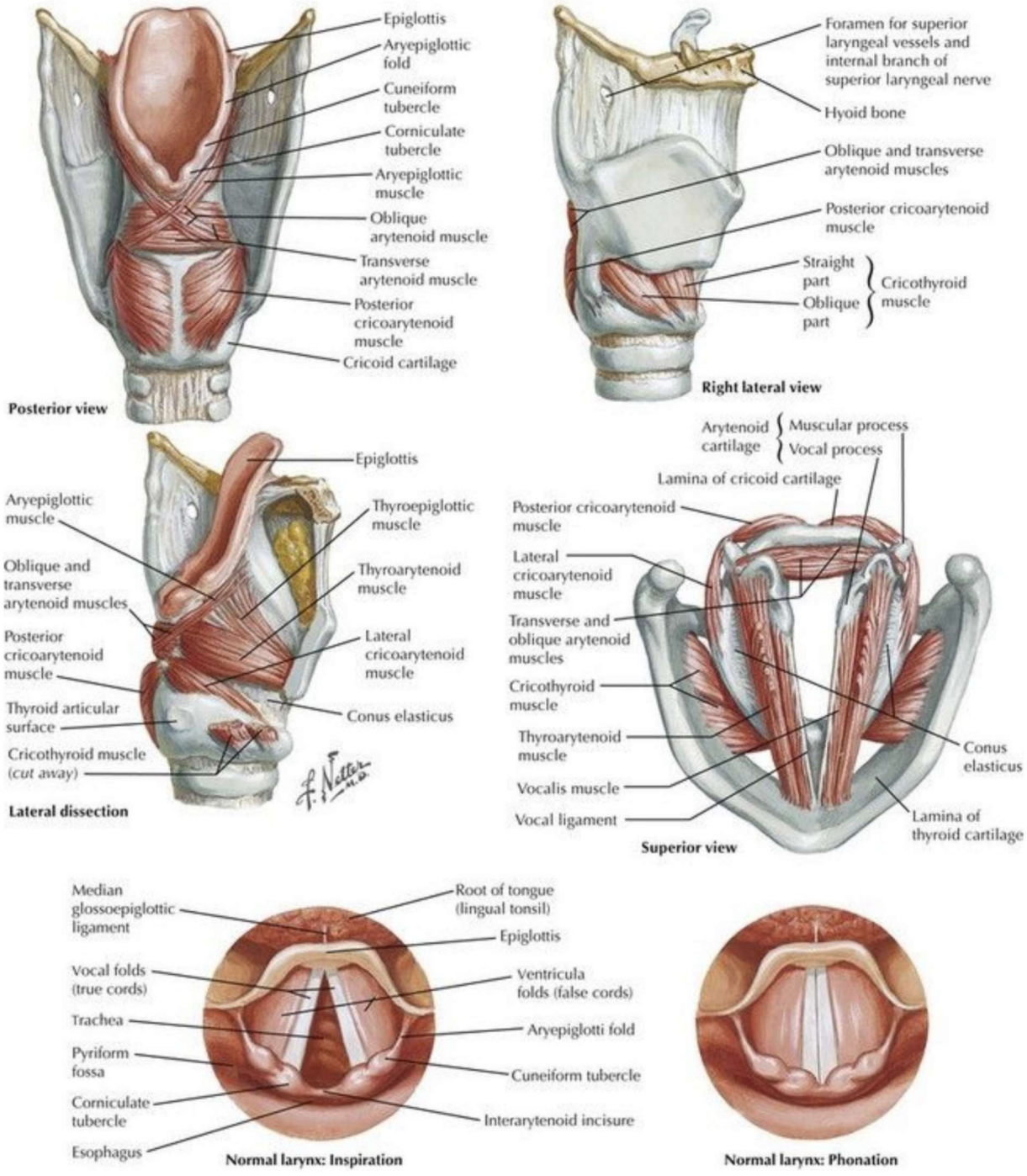


Figure 1.5: Anatomy of the larynx. Figure modified from Figure modified from Netter (2010).

folds to come apart. The vocal folds then are forced together again by a combination of elastic recoil forces and the aerodynamic Bernoulli effect (Titze, 1980; Van den Berg, 1958). The repetition of this glottic opening-closing cycle produces a periodic modulation of the passing airflow and this generates a sound wave. Thus, the sound *source* is the time-varying glottal airflow. The frequency of the sound wave (f_o) is the frequency of vibration of the vocal folds and Hn are its harmonics, integer multiples of f_o. f_o can be controlled and varied by acting on three mechanisms (Henrich Bernardoni, 2012): control on aerodynamic parameters via modifications of subglottal pressure and airflow rate, control on muscular parameters via contraction of the crico-thyroid, thyro-arytenoid and crico-arytenoid muscles (Fig. 1.5).

This explains how phonation relies on breath support to generate the correct subglottal pressure to meet the needs of human communicative tasks (MacLarnon & Hewitt, 1999). Phonation is such a complex process that the respiratory mechanisms of resting and even forced breathing are not sufficient. A necessary prerequisite for the production of phonation (and more broadly vocalization and speech in general) is a very fine control of the respiratory mechanisms and consequently of the subglottal pressure (MacLarnon & Hewitt, 1999), as well as a control of the balance between active and passive forces within the respiratory system (Huber & Stathopoulos, 2015).

1.1.3 Resonance and the Filter

As mentioned in the previous section, the auto-oscillation of the vocal folds produces a sound that contains many frequencies (Hn). When it leaves the laryngeal space, it is introduced into the supraglottic portion of the VT (the upper airways) where it is modulated, hence the designation of the supraglottic VT as the *filter*. Different harmonics are amplified or attenuated depending on the shape of the resonating cavities, i.e., the pharynx, the nasal cavity, and particularly the oral cavity. The resonant frequencies or formants (Fn) of the VT can be modified depending on the position of the mobile articulators and then radiated from the mouth.

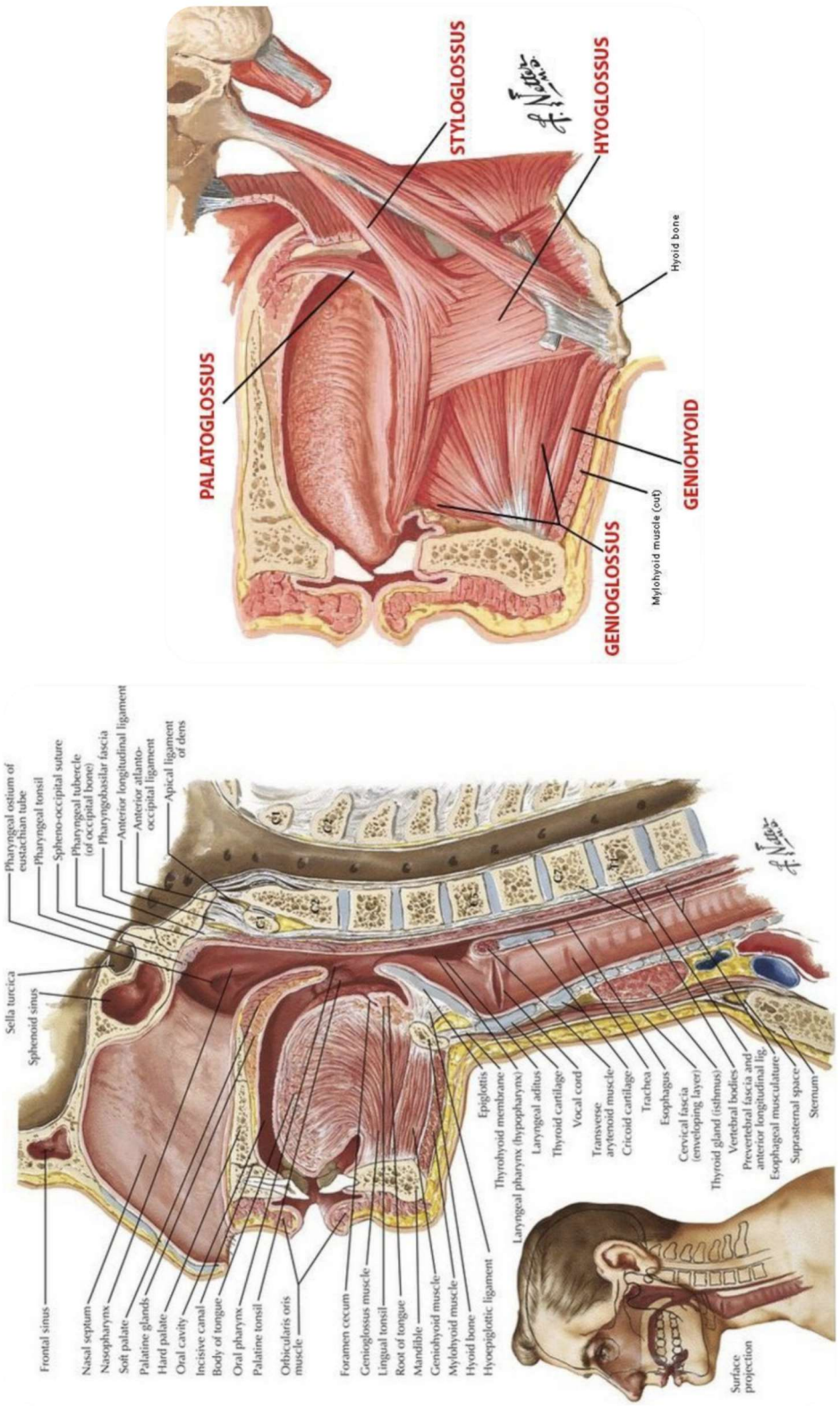


Figure 1.6: Anatomy of the vocal tract. Figure modified from Netter (2010).

The nasal cavity cannot change shape, but can be separated from the oro-pharyngeal cavity via the action of the velo-pharyngeal port. The shape of the pharynx can be controlled via the action of the pharyngeal muscles, that are divided into constrictor and elevator muscles depending on the orientation of their fibers. However, the presence of the tongue and its particular anatomy makes of the oral cavity the one that can be modified the most. The tongue is a muscular organ inside the oral cavity comprised of intrinsic and extrinsic muscles (Fig. 1.6). The extrinsic muscles of the tongue and their function are (Hixon et al., 2018):

- genioglossus: its superior fibers retract and depress the tongue apex, its middle fibers depress the tongue, and its inferior fibers protrude the tongue;
- hyoglossus: it depresses and retracts the tongue;
- styloglossus: it retracts and elevates the lateral aspects of the tongue;
- palatoglossus: it elevates the root of the tongue.

The intrinsic muscles of the tongue and their function are (Hixon et al., 2018):

- superior longitudinal: its retracts and broaden the tongue, it elevates the apex of the tongue;
- inferior longitudinal: its retracts and broaden the tongue, it lowers the apex of the tongue;
- transverse: it narrows and elongates the tongue;
- vertical: it broadens and elongate the tongue.

The length and aperture of the VT can also be modified at the lips via the action of the muscles of the buccolabial group (Hixon et al., 2018):

- the levator labii superioris, levator labii superioris alaeque nasi, risorius, levator anguli oris, zygomaticus major, zygomaticus minor muscles elevate and evert the upper lip;
- the depressor labii inferioris, depressor anguli oris, mentalis muscles depress and evert the lower lip;
- the orbicularis oris muscle closes the lips;
- the buccinator muscle compresses the cheek.

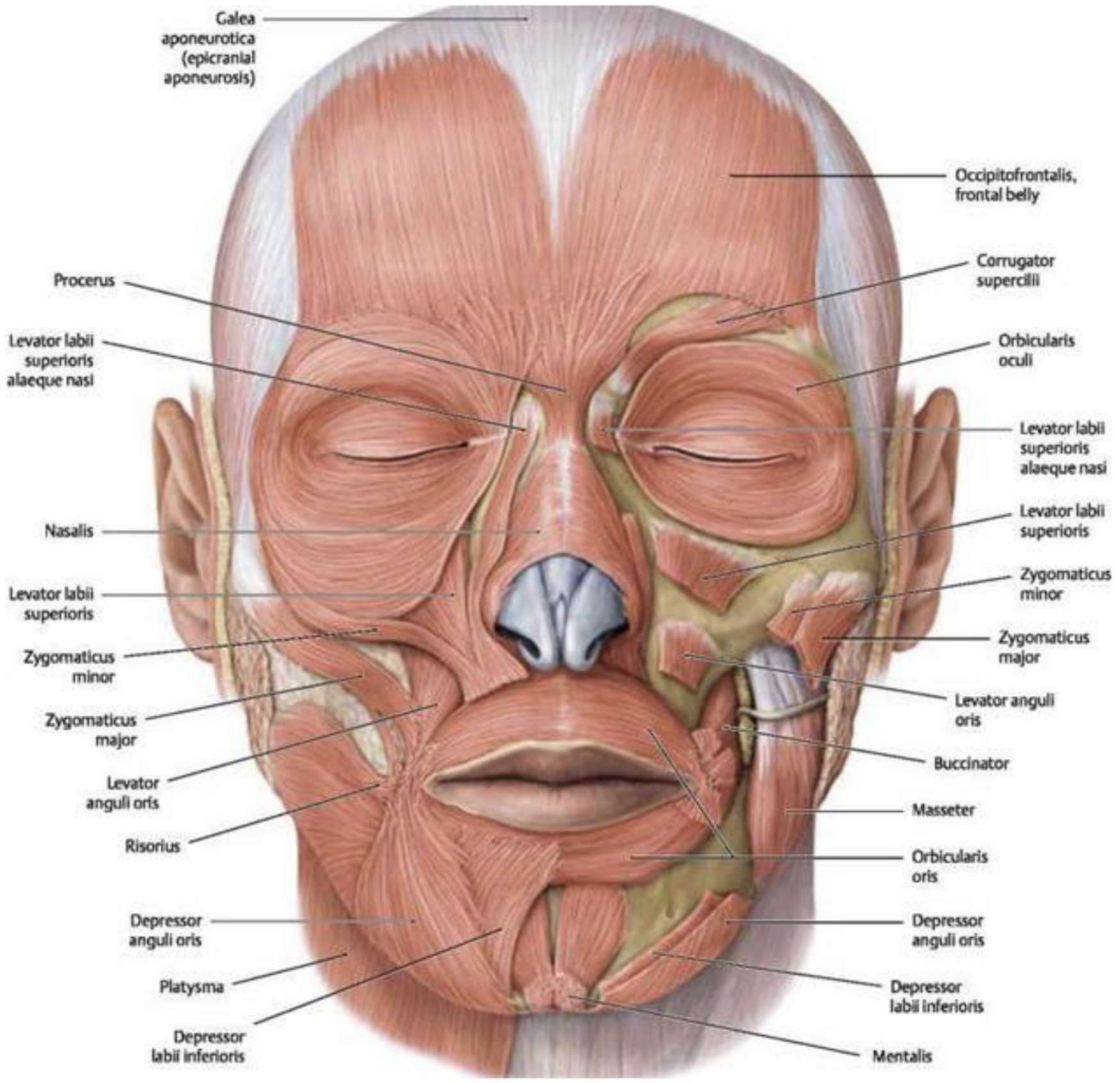


Figure 1.7: Overview of the facial muscles, including the buccolabial group. Figure from Atlas of Anatomy, Head and Neuroanatomy, Michael Schuenke.
From https://doctorlib.info/anatomy/atlas-anatomy/2.html

Further, the movements of the mandible or jaw also modify the volume of the oral cavity, under the action of (Hixon et al., 2018):

- the masseter muscle, a powerful closer that elevates the jaw;
- the temporalis muscle, another closer, it elevates and retracts the jaw;
- the internal pterygoid muscle, another closer;
- the external pterygoid muscle, it protrudes the jaw;

- the digastric muscle, composed of two bellies that depress the jaw and elevate the hyoid bone during chewing and swallowing;
- the mylohyoid and geniohyoid muscles, they both can either depress the jaw opening the mouth or elevate the hyoid bone.

1.2 Articulation and linguistic sound production

Of the multitude of sounds humans are capable of producing, only a relatively small subset has a linguistic function, and an even smaller subset enters the phonetic inventory of a given language. On a cognitive level, for a sound to be deemed linguistic, it needs to convey meaning: "the main point of language is to convey information" (Ladefoged & Disner, 2012). But what are the mechanisms that underlie speech production? Sounds may be selected for linguistic use based on physiological constraints on the VT. Ohala (1983) proposed that the speech production mechanism could be viewed as a device that converts muscular energy into acoustic energy. The chest walls, the larynx, and the tongue are compared to piston-like structures that generate direct-current pressure differences within the VT (Fig. 1.8). When these pressure differences equalize with atmospheric pressure, alternating-current pressure variations are created by the rapidly moving air. This produces sound. The chest walls, the larynx, and the tongue are called the initiators of the pressure change. The respective airstream so generated and the relative sound are called *pulmonic*, *glottalic*, and *velaric*. A compression would produce an *egressive* airstream, meaning that the pressure difference is equalized by expelling air from the VT to the outside. Conversely, a decompression would produce an *ingressive* airstream, meaning that the pressure difference is equalized by admitting air from the outside into the VT. From a physiological standpoint, 6 airstreams are possible, since in principle the three pistons can move in both directions. However, only 4 airstreams are attested to be used to produce linguistic sounds (Ladefoged & Maddieson, 1996; Ohala, 1983): pulmonic egressive, glottalic egressive and ingressive, and velaric ingressive. Pulmonic ingressive and velaric egressive may be used to produce para-linguistic sounds (Eklund, 2008; Ohala, 1983). However, the most commonly used airstream is by far pulmonic egressive (Helgason, 2014), which, as previously discussed, is essential to phonation and is also the only airstream that can sustain phonation in a modal register (Ohala, 1983). Generally, speech relies on phonation. The fact that most phonemes are also produced via pulmonic egressive airstream means that switching between different mechanisms is reduced. As to why there are no pulmonary ingressive phonemes in natural languages, Eklund (2008) points to the lack of control of ingressive airflow, the shape and function of the vocal folds, and the difficulty or impossibility of effectively reversing the action of the ventilatory system as the causes.

To summarize, the source-filter model Fant, 1981 gives a description of the acoustics

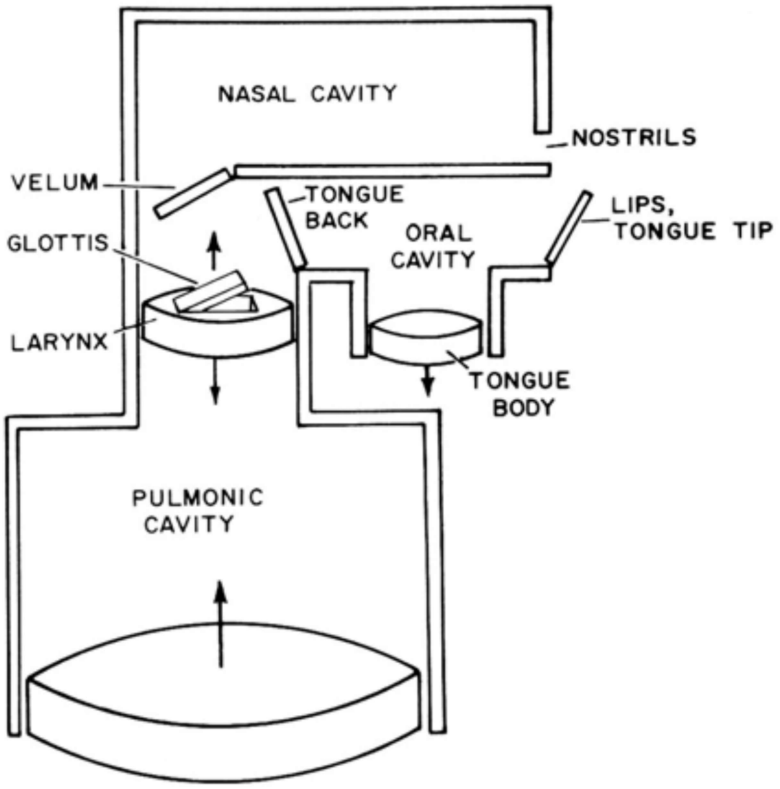


Figure 1.8: Ohala's schematic representation of the vocal tract as a device for the production of local pressure variations. Figure from Ohala, 1983.

of human speech production. In this model the principal sound sources are voicing and/or noise, both produced with pulmonic egressive airstream. The principal sound sourcings in speech are produced via a pulmonic egressive airstream mechanisms:

1. through adducted vocal folds: this sets the vocal folds into auto-oscillation and creates a source of vibration and thus voicing;

2. through constriction: this does not set the vocal folds into auto-oscillation, but produces turbulence and thus friction noise.

Other than an airstream, linguistic sounds are produced via the action of mobile articulators. The next section gives an overview of the most common articulatory mechanisms of speech based on the occlusion degree of the VT.

1.2.1 Articulatory mechanisms of speech

Ladefoged and Disner (2012) identify two principal areas of constraint to the selection of linguistic sounds: what our anatomy allows us to do and what our hearing allows us to distinguish. In particular, linguistic sounds would be selected based on one acoustic and two articulatory constraints, i.e., acoustic-auditory distinctiveness, ease of articulation, and gestural economy. Firstly from an acoustic standpoint, two sounds must have acoustic characteristics different enough for our hearing system to perceive and categorize them as distinct. Secondly, linguistic sounds generally do not require particularly complex gestures, and lastly, the same gesture can be used to produce multiple sounds (e.g., **t**, **d**, **n** are produced with a similar gesture of the tongue tip). The term articulatory gesture is rooted in the Articulatory Phonology framework (Browman & Goldstein, 1992; Browman & Goldstein, 1986), but here is used to indicate a pattern of movement for a family of linguistically equivalent articulations, and does not imply any formal theoretical value (Ladefoged & Maddieson, 1996).

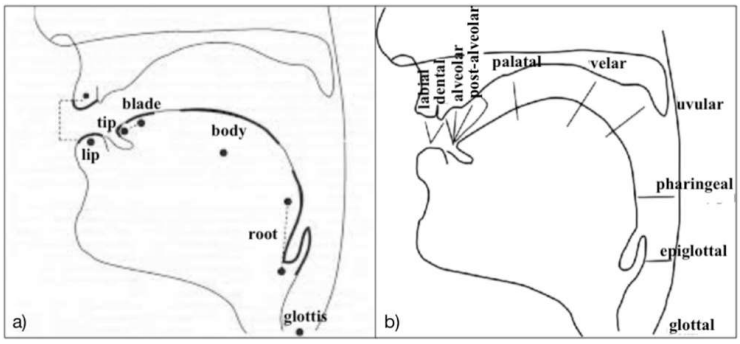


Figure 1.9: Mobile articulators and regions of articulation. Figure modified from Ladefoged & Maddieson, 1996.

These gestures are performed by mobile articulators that modify the shape of the VT by creating a constriction or a closure. Classically, mobile articulators (Fig. 1.9a) are the lips, the tongue, and the glottis (Ladefoged & Maddieson, 1996). More specifically, because of its anatomy and biomechanics, the tongue is divided in four regions of mobile articulation: tip, blade, body, and root. The movements of these articulators are aimed at more or less fixed areas of the VT (the articulatory targets), which determine the place (or more widely the region) of articulation of the phone (Fig. 1.9b).

However, more recent work has shown that the larynx is not only just a source of vocal fold vibration, but also a complex articulator. In the Laryngeal Articulator Model or LAM (Esling et al., 2019), the VT has two main complex articulators, one, the tongue, operates mostly in the Oral Vocal Tract, the other, the larynx, in the Laryngeal Vocal Tract (Fig.1.10a).

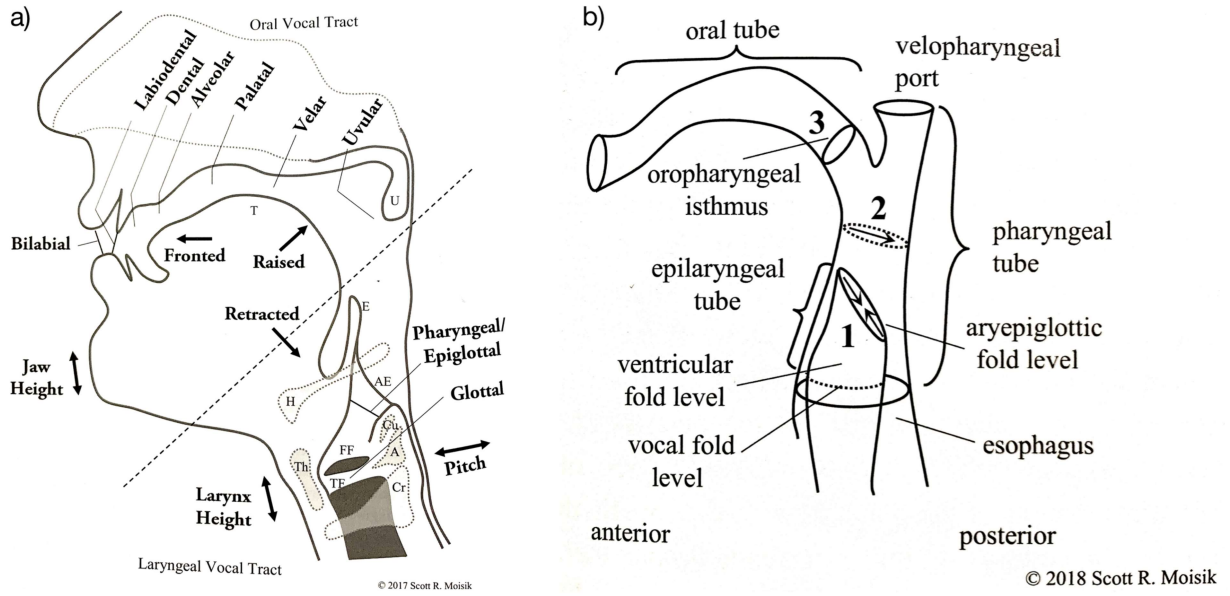


Figure 1.10: a) The Laryngeal Articulator Model, or the two-part vocal tract. T: tongue; U: uvula; E: epiglottis; H: hyoid bone; AE: aryepiglottic folds; Cu: cuneiform cartilage; A: arytenoid cartilage; Th: thyroid cartilage; FF: ventricular (false) folds; TF: vocal (true) folds; Cr: cricoid cartilage. b) The epilarynx as a tube-within-a-tube. Figure from Esling et al., 2019.

In this model, the combined action of the aryepiglottic constrictor mechanism, larynx raising, and tongue retraction is responsible for a large number of lower VT volume effects. This means that tongue root retraction and laryngeal raising are both related to laryngeal articulation. A more accurate description of the vibration sources of the VT does not restrain them solely at glottal level. The most efficient source of vibration is of course the auto-oscillation of the vocal folds. However, the ventricular folds and the aryepiglottic folds can be vibration sources as well. This redefines the active and passive articulators at laryngeal level: the locus of stricture is the upper border of the epilaryngeal tube (Fig. 1.10b), the active articulator is not the glottis, but rather the aryepiglottic folds, and the passive articulator is the epiglottis. Different levels of constriction result in different laryngeal configurations (Fig. 1.11): the first and less constricted is achieved with the complete closure of the vocal folds (configuration 5), complete closure of the ventricular vocal folds results in configuration 6, and the more constricted state (configuration 7) is reached via aryepiglottic closure.

Classically, two, or even three macro categories of linguistic sounds are described: consonants, produced via a certain degree of closure at a given point in the VT, vowels, generally voiced, articulated by the tongue via a non-localized narrowing of the VT, and semi-vowels or semi-consonants. Since this work focuses on consonantal sounds, we will briefly present

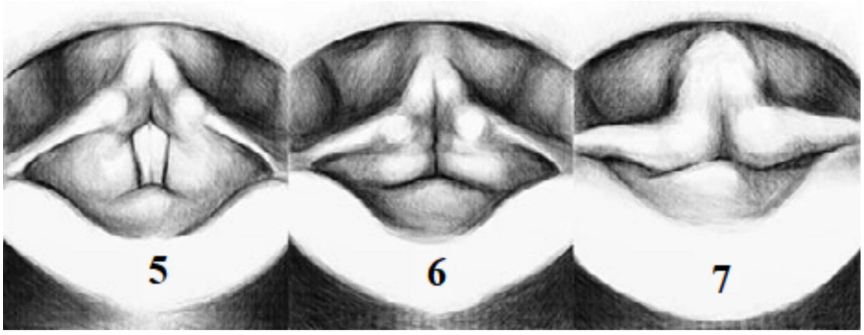


Figure 1.11: LAM Revised Open-Closed Continuum, configurations 5-7. Figure from Esling et al., 2019.

an articulatory description of speech consonants and somewhat neglect vowels. Suffice it to say, consonants can be considered as a means to begin or end a vowel:

"Consonants are nearly always movements at the beginning or end of a vowel. [...] they are best thought of as gestures of the tongue and lips [...]. Gestures are difficult to describe and it is easier to associate a consonant with [...] the target of the the gesture – the positions of the vocal organs that characterize the sound." (Ladefoged & Disner, 2012)

Three parameters are essential for the characterization and description of consonants:

1. vocal fold vibration: if vocal folds are adducted during the production of the consonant, the air coming from the lungs induces their oscillation and the resulting sound is said to be voiced; if the vocal folds remain abducted, there is no vibration and the sound is said to be voiceless;

2. place of articulation: it specifies where in the continuum of the VT the sound is produced, in other words, where the most significant modification of the VT takes place for the production of the sound in question;

3. manner of articulation: this indicates how the airstream is modified/managed to produce the target sound.

Within a language, speech sounds are distinguished from each other mainly on the basis of these three features. From the abstract point of view of phonology, contrastivity occurs between two or more phonemes. For example in French /p/, /t/ and /k/ are three different phonemes, distinguished according to the place of articulation: /p/ is a voiceless bilabial oral occlusive, /t/ is a voiceless alveolar oral occlusive and /k/ is a voiceless velar oral occlusive. From the phonetic point of view, the distinctiveness of phoneme realizations is categorical, between categories of sounds, i.e. small variations in the location of the articulatory target do not necessarily result in contrastive differences at the acoustic level.

However, some articulatory gestures that are quite different from each other can produce sounds that are acoustically very close, resulting in quantal articulations (Stevens, 1989). This means that, when producing speech, some variability is allowed in the trajectory of the articulator towards the consonantal target and possibly in the articulatory target itself, without the sound produced (especially the consonant) falling outside its phonological category.

A brief overview of the principal articulatory mechanisms of consonants follows. The mechanisms are presented based on a roughly decreasing constriction degree of the VT.

Oral and nasal occlusive consonants are articulated via the complete closure of the VT in correspondence of the place of articulation. The major difference between the two types of occlusives is the position of the velum: this structure is raised against the pharynx in the case of oral occlusives and lowered in the case of nasal occlusives. Therefore, when the velo-pharyngeal port is closed, air flows only through the oral cavity (oral occlusives). In contrast, when the velopharyngeal port is open, air flows through the nasal cavities (nasal occlusives). Both types of occlusives use a pulmonary airflow. In the case of **oral occlusives**, the closure of the VT at the place of articulation temporarily stops the airflow, resulting in an increase in oral pressure behind the occlusion and a sudden flow of air upon release of the occlusion. Acoustically, this closure produces a silence in the signal in the case of oral occlusives (or plosives) and a plosive sound in correspondence with the release of the occlusion.

In more detail, from an articulatory standpoint, three main phases can be identified in the articulation of oral occlusives:

1. *occlusion onset*: a mobile articulator (the lips or a region of the tongue, but also the glottis) moves to create a complete closure of the VT in the oral cavity or the larynx;

2. *occlusion hold*: the articulator remains in place while the pressure increases behind the closure. If the consonant is voiced, the vocal folds may stop vibrating towards the end of this phase;

3. *release*: the mobile articulator moves as the VT opens and the air is rapidly expelled. In the case of a voiced consonant, the vocal folds start vibrating again some time after the release (voice onset time or VOT).

From an acoustical standpoint, four events can be described (Calliope & Fant, 1989):

1. *silence*: it is consequence of the VT occlusion that prevents air from flowing and and causes an increase in intraoral pressure behind the closure;

2. *burst*: when the occlusion is suddenly released, the compressed air is rapidly expelled, causing an impulsive acoustic perturbation of variable intensity;

3. *release noise*: during the release phase of the occlusion, the articulators move to the articulation position of the next vowel. This movement is not instantaneous and depends on the speed of movement of the articulators. Therefore, in the point of the VT where the occlusion has taken place the air encounters a constriction that causes turbulence. The duration of the friction noise depends on the speed of movement of the articulators;

4. *formant transitions*: during this phase, formants are present, but are not stabilized, because the articulators are still moving towards the vowel target.

Relevant to this work, 6 oral occlusives are attested in French: /p, b, t, d, k, g/. They differ by the place of articulation, i.e., the point on the VT where closure takes place, and by voicing (/p, t, k/ are voiceless and /b, d, g/ are voiced). /p/ and /b/ are produced via the closure of the VT at the lips and are therefore called bilabial; /t/ and /d/ are two apico-alveolar occlusives where the occlusion is produced by the contact of the apex of the tongue with the alveolar part of the palate; finally, /k/ and /g/ are two dorso-velar occlusives and the occlusion is created by the contact of the dorsum of the tongue with the velum.

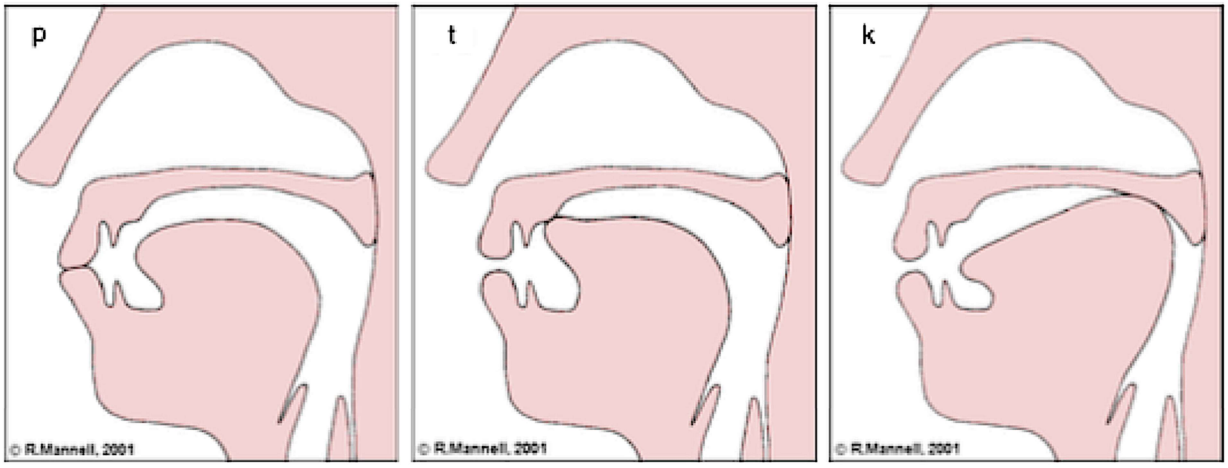


Figure 1.12: Schematic representation of the VT during the occlusion phase of the French voiceless oral occlusive consonants. Figure modified from http://clas.mq.edu.au/speech/phonetics/consonants/oral_stops.html

The voiceless oral occlusives of French (Fig. 1.12) are among the most widespread consonants in the world: according to Ladefoged and Disner, the phonemes /p, t, k/ figure in the phonologic inventory of about 98% of languages.

As mentioned above, during the articulation of **nasal occlusives** the velopharyngeal port is open. The occlusion of the VT occurs in the oral cavity prior to the velar opening (Ladefoged & Maddieson, 1996). Although the articulatory gesture in the oral vocal tract is similar between oral and nasal occlusives, the acoustic result is very different: in the case of nasals, there is no plosion, due to the fact that the airflow is not interrupted. In fact, the sound is continuous, as the air continues to flow through the nose throughout the maintenance of the oral occlusion, so that the intraoral pressure does not increase. Another difference occurs in the laryngeal vocal tract: nasals occlusives are almost always voiced, although different laryngeal configurations can be used (Ladefoged & Maddieson, 1996).

Occlusive consonants can also be produced using non-pulmonic airflow. This is the case of ejectives, implosives, and clicks.

Ejectives are produced by an egressive airflow of glottic origin. Two occlusions are created in the VT, one at the glottis, the other in the oral cavity. The air trapped between the two occlusions is compressed by the rapid rise of the closed glottis while the oral occlusion is maintained. As a result, the pressure behind the oral occlusion increases to about twice the usual pulmonary pressure (Ladefoged & Maddieson, 1996). The oral occlusion is then released and the compressed air can escape. Because of the higher supraglottic pressure than in the case of a pulmonic occlusive, in the case of ejectives, the burst is more intense from an acoustic standpoint. Ejective consonants are not widespread in the world's languages, however they are not rare either: they are present in about 18-20% of languages with the status of phonemes (Ladefoged & Disner, 2012; Ladefoged & Maddieson, 1996). Nonetheless, this type of mechanism seems to be employed in some phonetic contexts more and more commonly in languages like English and German (Simpson, 2014). Ladefoged and Disner (2012) point out that this type of consonant is not easy to produce. Furthermore, he observes that the velar ejective [k'] is easier to hear compared to the bilabial [p'] and the dental/alveolar [t'] and is more common.

In contrast to ejectives, injectives (or implosives) are produced using a non-pulmonic ingressive airflow. The closed larynx is rapidly lowered while the occlusion in the oral cavity is maintained. As a result of the increase in volume between the oral closure and the glottis, the air in this region may be more or less rarefied depending on the degree of closure of the vocal folds and the rate of descent of the larynx. The smaller the difference is between intraoral and extraoral pressure, the less intense the burst at the release of the occlusion. Injectives can be voiced and voiceless. Injectives are more difficult to produce than ejectives (Ladefoged & Disner, 2012) and are rather rare: about 10% of the world's languages have injectives (Ladefoged & Maddieson, 1996). The authors also observe that, despite voiced injectives showing multiple places of articulation, there is a tendency towards more anterior occlusions and bilabials are by far the most common.

Clicks are produced by an air mechanism classically called velar, always ingressive (Ladefoged & Maddieson, 1996). Two closures are created inside the oral cavity. The air present between the two closures is rarefied and a strong transient is produced at the moment of release of the most anterior occlusion. The rarefaction of the air inside the cavity is created through the action of the tongue. This constitutes the basis of the velar mechanism. Ladefoged and Maddieson point out that, although the names given to the different types of clicks refer primarily to the place of articulation, there is considerable variability in the articulatory gestures of clicks within the same category, depending on the language, but also among different speakers of the same language. In general, the production of an occlusive does not require a very high degree of articulatory precision:

"Making the articulatory closure for a stop involves simply moving one articulator so that it is held against another. It usually does not make much difference to the sound if the target position, which is above the upper surface of the vocal tract, is a few millimeters higher so that there is a tight closure, or lower so that the closure is formed more gently. A stop closure will produce more or less the same sound as long as it is complete, irrespective of whether there is firm or light articulatory contact." (Ladefoged & Maddieson, 1996)

Trill consonants are produced through the vibration of a mobile articulator against a surface of the VT: the articulator is positioned close to the surface and, when an airflow of a specific intensity passes through the aperture created by this configuration, the mobile articulator is pushed by aerodynamic forces against and away from the surface several times, causing a repetitive pattern of closing and opening of the airflow channel. Ladefoged and Maddieson (1996) observe that this movement is similar to the vibration of the vocal folds during voicing. In both cases, the movement that creates the vibration is not muscular, but aerodynamic in origin. Therefore, the articulatory gesture of bringing the mobile articulator close to the surface of the VT must be extremely precise for the trill to occur and even small deviations cause the trill to fail. In principle, it is possible to prolong a trill as long as the aerodynamic conditions remain favorable. In fact, in the languages of the world that use trills, only 2-5 periods are produced, a little more in geminate productions. As Ladefoged and Maddieson (1996) point out, trills are easier to produce if the vibrating articulator has a low mass. This results in the most common trills being produced by the tip of the tongue vibrating against the dental/alveolar region and the uvula vibrating against the back of the tongue.

Flap or **tap** consonants are produced similarly to trills. In contrast, only a single very brief closure occurs between the mobile articulator and the VT surface. Flaps are usually apical.

Constrictive or **fricative** consonants are articulated via a strong narrowing of the VT in correspondence with the place of articulation. From an aerodynamic point of view, this narrowing produces turbulence in the airflow. This results in a friction noise in the

acoustic signal, hence the name fricatives. There are several types of fricatives, depending on the position where the constriction is created in relation to the medio-sagittal plane as well as the location where the turbulence is produced. Thus, when the fricative sound is created in the medio-sagittal plane, the fricative is called central and when it occurs laterally to the articulator (e.g., the air flows on one or both sides of the tongue), the fricative is called lateral. Generally, fricative sounds are the result of turbulence produced at the narrowing of the VT. However, in the case of sibilants, the narrowing of the VT is used to accelerate the outgoing air and create a jet of air that impacts against the edge of an obstruction such as the teeth. The most common fricatives in the dental, alveolar and post-alveolar regions are sibilants. Generally, fricative consonants are produced with egressive pulmonic airstream. Nevertheless, in some languages some fricatives are produced via glottalic airstream, namely in the case of ejective affricates. Fricatives are consonants that can be articulated throughout the VT. They are common in the world's languages, but somewhat less common than occlusives or nasals (Ladefoged & Disner, 2012). Unlike occlusives, the articulation of fricative consonants requires a very high degree of articulatory precision, due to aerodynamic constraints: for a turbulent flow to be produced, it is imperative that the VT be configured in a very precise shape and that this shape be maintained constant over a certain period of time, especially in the case of sibilants (Ladefoged & Maddieson, 1996).

Two types of approximant consonants exist: central approximants and lateral approximants. In general, they are produced via a pulmonic egressive airflow. **Lateral approximants** are articulated through an occlusion somewhere along the mid-sagittal line of the VT. Because the occlusion is incomplete laterally, air can escape from one or both sides of the occlusion. Most of the lateral approximants of the world's languages are produced with occlusion at the dental/alveolar region (Ladefoged & Maddieson, 1996). **Central approximants**, or simply approximants (sometimes called **semi-vowels**) are vowel-like segments that act as consonants. They are articulated via a slight narrowing of the VT in correspondence with the place of articulation. Such narrowing is still more pronounced for the approximant than for its corresponding vowel. Some approximants are common in the languages of the world: 85% have the palatal approximant [j] and 76% the labio-velar approximant [ɰ], but other approximants are much rarer, appearing in only 2% of the languages. This is the case for the labio-palatal approximant of French (Ladefoged & Maddieson, 1996).

In conclusion, the main articulatory mechanisms of speech sounds of the world's languages and the degree of articulatory precision required have been briefly described. In sum, fricatives and trills require a very high degree of articulatory precision, whereas in the case of occlusives a lower degree of precision is sufficient. Furthermore, it has been shown that most of the sounds of the world languages are produced by exploiting the egressive airstream mechanism.

1.2.2 Breathing behavior of speech

As previously mentioned, during phonation the respiratory system ensures the correct subglottal pressure allowing the closed vocal folds to vibrate. Therefore, phonation usually takes place during the expiratory phase of the breathing cycle. However, during phonation, the breathing cycle (called a breath group when phonation is present) adapts to the needs of phonation and the asymmetry of the cycle is increased: the inspiratory phase becomes short and fast, and the expiratory phase becomes longer and slower in terms of the flow rate of exhaled air. The linguistic message is transmitted through speech. This generally occurs during the expiratory phase. In order not to be too fragmented, the speaker must be able to produce a certain number of words before pausing their speech flow to take in air. Thus, the inspiratory phase cannot occur at any time, but must be adapted to the needs and structure of the language. Therefore, air intakes must be quick, while mobilizing the necessary breath volumes and occur at linguistically relevant times (Huber & Stathopoulos, 2015; MacLarnon & Hewitt, 1999) when they do not interfere with the message conveyed by the speaker (e.g., between one utterance and the next). The mechanisms of resting breathing cannot meet the needs of phonation and speech production in general, and air volumes must be managed differently. The elastic recoil forces alone cannot produce the subglottal pressures necessary for phonation. Therefore, the expiratory phase becomes an active phase. Conventionally, the breath group is subdivided into four phases, depending on the muscular effort involved (Titze & Martin, 1998). Three phases take place during expiration, whereas inspiration accounts only for one phase. During the initial phase of expiration, adequate pressure can be achieved through the elastic recoil forces of the lungs and rib cage. In fact, if inspiration is deep enough, elastic recoil can generate excessive pressure, which can be controlled by briefly prolonging the contraction of the diaphragm in expiration (Leanderson et al., 1984), but also via the external intercostal muscles, whose action controls the descent of the ribs and counterbalances the elastic recoil by slowing down the flow of expired air (Draper et al., 1959). During the second phase of expiration, the internal intercostal muscles contract and reduce thoracic volume until elastic recoil is exhausted. At this point, the third phase begins where the abdominal muscles contract to produce the major contribution to lung pressure (Hoit et al., 1988; Watson et al., 1989). In addition, the back muscles may also be recruited to compress the thorax (Hixon, 1973). At this point, the rib cage and lungs are so compressed that they provide negative recoil (and therefore negative pressure) that must be overcome by more abdominal effort. During the inspiratory phase, the abdominal muscles (rectus abdominis, external oblique, internal oblique, and transverse abdominis) and internal intercostal muscles must promptly relax to allow for rapid expansion of the lungs, achieved through contraction of the external intercostal muscles and diaphragm, aided by negative recoil of the compressed lungs and rib cage. Thus, when speaking, breathing is actively managed by the respiratory muscles, both during the inspiratory phase, and especially during the expiratory phase. Moreover,

the respiratory muscles not only cause the volume changes necessary to mobilize the air, but also take care of controlling the rate of such changes and consequently the flow of air ventilated (Huber & Stathopoulos, 2015). Such muscle activation requires a higher energy cost compared to resting breathing (Huber & Stathopoulos, 2015), although it has been shown that individuals tend to stay as close to the functional residual capacity as possible (Cerny & Burton, 2001). From an evolutionary perspective, the motor system has had to undergo reorganization to accommodate the respiratory needs of speech without compromising O_2 and CO_2 gas exchange (Titze & Martin, 1998).

Linguistic sound production occurs during exhalation. However, research conducted by Eklund (2008) has shown how sound production often occurs during inspiration in what he calls *pulmonic ingressive phonation*. The author states that this type of phonation is widespread throughout the world, regardless of the language family, and apparently always performs the same communicative functions. Therefore, he hypothesizes that pulmonic ingressive phonation might represent a universal linguistic phenomenon, rather than a "strongly typologically marked" form of phonation. However, pulmonic ingressive phonation would not have a strictly linguistic function. Rather, it would fulfil para-linguistic functions such as a positive feedback marker produced in a more or less unconscious way, or vocal dissimulation (e.g., the case of ventriloquism) produced in an intentional way, or the expression of emotions (e.g., surprise, fear, etc.).

1.3 Non-linguistic sound production

When we talk, we do not restrain our communication to linguistic material. Non-linguistic sounds are often integrated to linguistic production for various communicative purposes or to modify the meaning of what is said and convey emotion (Eklund, 2008). In certain languages including English and French, it is common to express emotion-related meaning such as disapproval, surprise, pain, doubt, etc., very effectively with short vocalizations, without recurring to verbal explanations. These sounds are not subject to strong linguistic constraints and therefore, in theory, the full potential of human voice production can be exploited in the production of non-linguistic sounds. Indeed, Eklund (2008) reports large paralinguistic use of clicks in several languages, namely French and English, that do not include clicks in their phonology. For instance, *tsk-tsk* or *tut-tut* are used to express disapproval in English; a dental click can signify denial in French, or a bilabial click often represents a kiss and is sometimes accompanied by phonation for emphasis ([mːʘwa]). Further, the use of pulmonic ingressive phonation is widespread: doubt, delight, hesitation, indifference, pain, satisfaction can all be expressed via pulmonic ingressive phonation. In certain languages such as French and Swedish whole words can sometimes be produced via pulmonic ingressive phonation: such is the case of the word *yes* in both languages.

Other than emotion, non-linguistic sounds can be used to designate and describe animated and inanimate objects of the world, without words. Sound imitations are often integrated to speech production and constitute an effective way to communicate. In fact, vocal imitations of sounds have proven to have some degree of convention and to be more effective than verbal descriptions (Lemaitre et al., 2016; Lemaitre & Rocchesso, 2014). Research has shown that the effectiveness of vocal imitations does not rely on accurate reproduction of the sound, but rather on emphasis on a few important acoustic features, such as pitch, tempo, sharpness, and onset (Lemaitre et al., 2016). However, these features may differ depending on the category of the sound being imitated (Lemaitre et al., 2016). As Friberg et al. (2018) point out, the sounds used to imitate environmental sounds or animals have some similarities with speech sounds, but also go beyond speech-like characteristics. Acoustic features such as jitter, shimmer, cepstral peak prominence, and noise-to-harmonic ratio have been developed to automatically quantify voice characteristics (see Friberg et al., 2018) in speech production. However, in sound imitations humans can utilize a far wider range of articulations than are used to make phonological distinctions in languages. Imitators can utilize mechanisms that are typologically rare and considered "difficult" (Helgason, 2014). A machine learning (ML) approach has proven successful in automatically predict the articulatory category of vocal imitations sounds based on recorded audio (Friberg et al., 2018). This approach is based on identifying a range of features for each articulatory category and subsequently predict the actual articulatory mechanism of production of the sound. Indeed, a framework based on articulatory and aerodynamic mechanisms, rather than acoustic, seems very promising for the characterization of non-linguistic sounds. Based on Pike's approach (Pike, 1943), Helgason (2014) presents a classification of sound production based on 3 basic source types (myoelastic, turbulent, whistled) intersecting with 6 basic sound initiation mechanisms (pulmonic, glottalic, velaric, both ingressive and egressive). Percussive sounds form a category on their own, being both an initiation mechanism and a source type. This system is based on the fact that the common way of humanly producing sounds is to move air inside the VT and presenting it with an obstacle (see section 1.2.1). The air can be set in motion by different structures (lungs, larynx, tongue): this is the initiator. The obstacle is where the sound is produced, that indicates the source type. Turbulent sources produce fricative sounds; myoelastic sources produce sound via the oscillation of elastic tissue that can be perceived as a tone; whistle sources produce whistle sounds by directing a jet of air against an obstacle; and percussive initiation produces sound via impact of two solids. Helgason (2014) was able to successfully characterize sound imitations using this classification. In his corpus, he was able to identify the use of turbulent sources combined with pulmonic egressive, glottalic egressive and ingressive, and velaric ingressive airstreams. Laryngeal myoelastic source coupled with a pulmonic egressive initiation mechanism (i.e., voicing) was the most frequently observed, as in speech. However, other places where vibration was produced are the aryepiglottic folds, the lateral edges of the tongue, and the lips. Different tension at the lips was used to produce

different sounds. Myoelastic sources were also coupled with pulmonic ingressive initiation (i.e., ingressive phonation). Whistled sources were observed in conjunction with pulmonic egressive (labial) and velaric ingressive initiation mechanisms. One sublaminal percussive sound was described. This framework seems therefore very convenient, in that it includes all the possible mechanisms of sound production of the VT, including speech sounds, but can easily be applied to more exotic non-linguistic sounds.

1.4 Singing

Voice can be used for artistic purposes. Linguistic sound production (i.e., lyrics) can be merged with music, which results in singing. This shift from linguistic to non-linguistic production produces modifications at various levels. As speech, singing conveys linguistic meaning through words and sentences, but in this case the most important constraints are of musical nature. Sound production is therefore adapted to take into account the different constraints. Of course, every human being sings. Some sing everyday: under the shower, while commuting, in the office... However, no one becomes a professional singer overnight. This is especially true for Western opera singing or classical singing in general. 6 to 8 years of intensive training are common before a singer can access a professional career. During this extended training, an aspiring singer must learn to adapt their normal articulatory behavior to an acoustically more demanding sound production that has to overcome an orchestra and has to have a particular aesthetic quality as well (Austin, 1997). Further, adaptations are so extensive that, without prior exposition to the lyrics of an aria, the text becomes largely unintelligible. In singing in general, but particularly in classical singing, articulatory adaptations occur for acoustic reasons, and appear to happen mostly on vowels. In speech, f_o modulations (prosody) are essential to convey lexical meaning (e.g., in tonal languages), grammatical meaning (e.g., interrogative sentence as opposed to declarative sentence), or emotional meaning. However, they occur over a smaller range than singing, and rarely if ever span over the whole vocal range of physiologically possible frequencies of vibration of the vocal folds. In singing, modulations of f_o produce melody, and this can take place over all the vocal range, covering multiple octaves. In fact, one aspect of classical singing training aims at extending the singer's vocal range over lower and higher f_o. Furthermore, singing requires the production of a precise f_o at any given time. This means that the mechanisms that regulate the frequency of vibration of the vocal folds and their modulation must be finely controlled. When singing vowels, f_o higher than speech means that f_o can exceed the normal F1 for a given vowel. This is disadvantageous from an acoustic perspective. Singers have been shown to rely on adjustments of the VT through articulatory manoeuvres such as jaw lowering to avoid this situation (Sundberg & Skoog, 1997). Other articulatory adjustments such as reducing tongue dorsum height and increased lip opening were shown to fit the same purpose and to gain in acoustic intensity

(Sundberg, 1987). In fact, jaw opening seems to be so paramount for classical singing that even pupils at the beginning of their training have already integrated this manoeuvre in their technique (Austin, 1997). Articulatory adaptations of the tongue position result in centralized vowels (Dromey et al., 2011): front vowels are pushed backwards and back vowels are pushed forward. This results in a blending of the acoustic signature of the vowel to such an extent that they become indistinguishable from one another (Hollien et al., 2000). Therefore, it becomes almost impossible for the listener to correctly perceive words and access their meaning (Gregg & Scherer, 2006).

The articulatory adaptations are not the same over the vocal range. Rather, they seem to differ with respect to pitch, but also loudness (Echternach et al., 2016). For instance, the vertical larynx position increases with increasing f_o, but it is lower when greater loudness are concerned. Lip opening and pharynx width correlate with sound pressure level more than with pitch. Variable larynx height is also observed for speech vowels, but in this case the differences are related to the position of the tongue in front as opposed to back vowels. In singing, the larynx seem to assume more similar heights among different vowels, and a lower average position than in speech (Sundberg, 1969).

A lower position of the larynx together with increased lip protrusion modify the length of the VT. The longer VT produces a shift in formant frequencies, especially F1 for back vowels and F2 for front vowels. Further, larynx lowering has been related to the appearance of the so called "singing formant" in professional male singers (Sundberg, 1974), i.e. a region of high spectral energy near 3 kHz characteristic of vowel sounds. When vowels preceded or follow nasal consonants, the velopharyngeal port is increasingly narrowed as pitch increases (Austin, 1997). The enlargement of the pharyngeal cavity supports the assumption that vowels are articulated without nasality also in singing (Sundberg, 1969).

However, articulatory configurations seem to be different depending on the style. Indeed, the acoustic characteristics of the sound production differ among different singing styles. For instance, in classical singing usually the lower harmonics are boosted, whereas in popular music it is rather the higher harmonics.

In addition to adaptations at the articulatory level, extremely fine control of ventilation is required. While most individuals can sing spontaneously, expert production requires years of training to acquire the necessary control over the phonatory organs, but also over the breathing mechanisms. In fact, singing long phrases in continuous phonation may require almost the entire expiratory reserve volume (Titze & Martin, 1998, p.67). Therefore, active breath management is fundamental in order sustain a sung phonation. Moreover, the singer cannot stop at any time to take in air, since they must respect the structure of the music. The control of the breathing mechanisms is obtained through the action of the respiratory muscles, which are used in a different way than in speech. Professional classical singers often advocate active control of the abdomen in order to achieve better singing

performance. More efficient vocalization with better projection requires greater abdominal support, which professional singers achieve through greater activation of the abdominal muscles (Thorpe et al., 2001). The abdomen is used to prevent a too rapid rise of the diaphragm. This would result in a more efficient generation of subglottal pressure during phonation. Indeed, classical singers dissociate thoracic and abdominal breathing kinematics, use a greater abdominal contribution to respiratory volume compared to inexperienced individuals, but also perform inward prephonatory abdominal movements that increase intra-abdominal pressure, presumably in an effort to increase the pressure-generating capacity of the expiratory muscles of the rib cage (Salomoni et al., 2016). In addition, expiratory muscles such as the upper trapezius, internal intercostals, oblique and anterior abdominals, sternocleidomastoid, and scalene are activated (Pettersen, 2005). However, although a breathing strategy adapted to singing, particularly lyrical singing, is essential, it does not seem to be uniform in all singers (Thomasson & Sundberg, 2001). On the other hand, the ventilatory strategy of the same singer is very reproducible (Thomasson & Sundberg, 2001).

In conclusion, the breathing behavior must be modified from speech to meet the needs of the singing production. The use of inappropriate breathing technique and airflow generation may result in inadequate phonation support. As a result, vocal effort and muscle tension may increase.

Chapter 2

Human Beatboxing

Contents

2.1 A rapidly evolving vocal art

Human Beatboxing (HBB) is a relatively young, yet extremely diverse urban vocal art belonging to the Hip-Hop culture. It originated in the USA in the late '70s, early '80s with the aim of reproducing the sounds of the electronic beat boxes (hence the name Human Beatboxing), in particular the Roland TR-808 (Fig. 2.1). Vocal instrument mimicry allowed young and disadvantaged people to do music, both the vocals and the instrumentals, with nothing more than their voice, or rather their body.

Figure 2.1: A beat box Roland TR-808.

Hip-Hop music is essentially characterized by two elements: vocals or the melodic line,

provided by the MC[1] or rapper, and a rhythmic base or line, usually generated by a beat box. Here is where HBB came into play: the expensive electronic beat boxes were replaced by the human voice[2] and the beatboxer provided the rhythmic base to accompany the vocals. The reader may see why the foundation core of HBB sounds is made up of mostly sounds imitating percussion instruments. However, the art rapidly evolved and expanded to the imitation of the electronic effects generated by the synthesizers used by the DJs[3]. More recently, HBB has become more complex and diverse (Ojamaa & Ross, 2009): not only beatboxers imitate the sounds produced by instruments, but more broadly they experiment with their voice in order to come up with new and innovative sounds never heard before, so that their repertoire is in constant development. Nowadays, beatboxers also have more resources and this has allowed them to reintroduce technological tools into their musical practice. Loopstations allow for live recording of HBB effects and looping them by layering multiple sound tracks. Thus, a beatboxer can recreate or create an entire song in front of an audience, combining beatboxing and looping. Or, the beatboxer can record their HBB sounds and have them played by the instruments he imitates instead of their original sound. In this respect, HBB is a very prolific environment for the experimentation and evolution of human sound production. From a scientific perspective, this is very captivating, in that the understanding of the mechanisms of sound production exploited in HBB could shed light on multiple domains of research, such as Phonetics and Phonology, Physiology, Health Sciences, etc.

HBB as a vocal art has always been extremely personal in many ways. First of all, no formal and shared pedagogy exists to this day: often neophytes self-educate to HBB, by trying to reproduce instrumental and environmental sounds, or by imitating other more experienced beatboxers. Second, no two repertoires are alike: while the acoustic result of basic HBB sounds may be similar among all beatboxers, each beatboxer has their signature sounds and techniques, to differentiate themselves from the others. Nevertheless, in recent years the international HBB community has shown growing interest in the exchange of techniques and sonorities, and national and international bootcamps[4] and competitions are organized on a regular basis. In this scenario, Internet has become an important tool for the exchange of knowledge and techniques. On social media platforms such as Instagram and YouTube beatboxers post their tutorials, reactions, and original creations. They can exchange and interact with other beatboxers all over the world. HBB being predominantly oral, to this day there is no satisfactory writing system that would facilitate exchanges and allow the creation of a long-lasting record of their vocal productions. Efforts have been ongoing for decades: Tyte and Splinter developed a method for representing sounds

[1]Master of ceremony, a vocal artist who creates vocals.

[2]We use voice in a broad sense, indicating the whole range of sound production the phonation organs can produce.

[3]Disc jockey, a person who plays recorded music for an audience.

[4]Intensive training sessions that can take place over several days.

and beat patterns[5], the Standard Beatbox Notation (SBN) system. This system uses characters from the standard English alphabet and combines "typography and phonetics to use letter shapes as an image of their sounds" (TyTe & SPLINTER, 2002). From a scientific standpoint, Proctor et al., 2013 have proposed a notation system that uses a combination of the International Phonetic Alphabet (IPA) and the "standard percussion notation". As Contesse (n.d.) points out, these early efforts based on linguistic notations led to notation systems that were cumbersome and not entirely adapted to represent the very rich and constantly evolving sound universe of HBB, which also includes a multitude of sounds whose root is not linguistic. They implemented Vocal Grammatics, a modular system of about 30 signs (called "glyphs") representing the information necessary to produce the desired sounds. This system is based on articulatory phonetics and each "gliph" provides visual and intuitive information on the location (i.e. place of articulation) and articulation (i.e. manner of articulation) of the sound.

2.2 Scientific characterization of HBB production

HBB being a recent art form, very few scientific studies have been conducted on this topic. Nevertheless, the very nature of HBB, "hybrid: vocal, but not linguistic, musical, but not instrumental" , makes it interesting for research in many fields, from music information retrieval (MIR) (Kapur et al., 2004; Sinyor et al., 2005) to automatic recognition of HBB sounds (Picart et al., 2015). Relevant to our work, HBB has aroused the interest of experimental phonetic scholars, since it offers the possibility to explore the potential of VT functioning (Blaylock et al., 2017), in that "the articulatory phonetic performance details of HBB go beyond typical combinations and range of sounds found in the speech of most languages" (De Torcy et al., 2014, p.10), to the point of using extreme and rare configurations (De Torcy et al., 2014; Proctor et al., 2013; Sapthavee et al., 2014) and places and articulation mechanisms that have not yet been attested in speech (Blaylock et al., 2017; Proctor et al., 2013). However, Proctor and colleagues observe 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" (2013, p.1050). This observation is nevertheless strongly contested by Blaylock et al., 2017, who point to the widespread use of articulatory patterns and airstreams within HBB that are not attested in speech, such as for instance ingressive retroflex trills, ingressive lateral bilabial trills, lingual egressive and pulmonic ingressive airstreams. Presumably, a vision such as that of Proctor et al., 2013 is adapted to the HBB of the beginnings and to the basic sounds of this vocal art, while a vision such as that of Blaylock et al., 2017 better reflects the practice of HBB nowadays. In any case, the studies conducted by De Torcy

[5]A combination of beatboxed sounds.

et al., 2014; Sapthavee et al., 2014 show a very interesting scenario: beatboxers learn to exploit their articulatory structures to the maximum, to use them rather independently of each other, to control them via a high developed proprioceptive feedback and to set up mechanisms to protect against glottal injuries. In addition, the articulatory , acoustic and aerodynamic characteristics of HBB sounds appear to be more pronounced than speech sounds, suggesting that the study of HBB could provide valuable contribution to achieving more complete functional models of articulation, articulatory and phonatory coordination, and speech control. Further, such advances in knowledge of the functioning of the vocal organs could be exploited in speech therapy in many areas, such as dysphonia therapy, post-laryngectomy rehabilitation, myofunctional therapy, etc. By virtue of the skilful use of voice production mechanisms and ludic nature, HBB has made its way into Speech Therapy and some clinical studies have already shown promising results. A therapeutic tool has been designed using occlusive consonants to treat dysphonia in patients with unilateral laryngeal immobility (Poupineau, 2018). Two recent clinical studies have introduced beatbox into a voice (Le Meillour, 2019) and swallowing (Navarro, 2019) rehabilitation protocol in Parkinson's patients. It is used as a playful therapeutic tool for speech and language development disorders in children (Mignot, 2018). A 6-week therapy program based on *beatalk* provided larger gains in articulation accuracy and voice measure than traditional therapy in adults with intellectual disability, probably due to the more 'fun' and hence engaging nature of the activities (Icht, 2019). It has been shown that laryngectomized patients can beatbox (Himonides et al., 2018). However, it is our opinion that better understanding of the production mechanisms involved in HBB is advisable to adequately integrate HBB in speech therapy protocols and avoid risks of muscular straining and unhealthy vocal behavior in voice patients. Recent studies have shown that prolonged (>18 months) HBB practice may have an adverse effect on vocal musculature, and modify jitter, shimmer, and harmonic-to-noise ratio values (Verma et al., 2019). Beatboxers also reported vocal complaints such as vocal fatigue during and after long performances, breathing difficulties during and after performance, and muscle tension while beatboxing, as well as the resort to non-vocal habits such as on-time food intake and relaxation techniques prior to performance. However, despite reporting higher scores of vocal fatigue than non-singers, beatboxers seem to have lower scores compared to untrained singers, suggesting that they develop a better management of the laryngeal musculature, hence protecting the vocal folds from strain (Dodderi et al., 2020).

From a scientific perspective, the human-voice sound production that is HBB is captivating, because beatboxers explore all the possibilities of their vocal instrument unrestrained by style or language. However, the existing literature on HBB comprises only a few published studies. The earliest works dealt with automatic recognition and classification of basic HBB sounds based on acoustic data. Kapur et al. (2004) exploited acoustic features of some HBB sounds and rhythmic information for a new approach on Music Information Retrieval (MIR). Sinyor et al. (2005) tested the use of the Autonomous

Classification Engine (ACE) for classifying some basic HBB percussion sounds. More recently, Picart et al. (2015) and Evain et al. (2020) investigated the automatic recognition of pre-recorded HBB drum and instrument sounds. Pillot-Loiseau et al. (2020) have provided physiological details on the production of some HBB sounds. They also compared the acoustic characteristics of HBB imitations to the acoustic characteristics of the homologous sounds produced by the actual instruments. They found that attack, harmonic distribution, spectral envelopes, waveforms, and energy distribution were very well reproduced. Other studies have investigated the production mechanisms of HBB sounds. De Torcy et al. (2014) and Sapthavee et al. (2014) conducted endoscopic investigations on the laryngeal structures involved and the overall behavior of the larynx during beatboxing. They showed very active laryngopharyngeal dynamics and a dissociated mobilization of the laryngopharyngeal structures. These authors pointed out the use of extreme articulatory configurations in the laryngopharynx region, a piston-like action of the closed glottis that accompanies the production of some plosive sounds (De Torcy et al., 2014) as well as articulatory behaviors that can protect against glottal injury (Sapthavee et al., 2014). Some studies have explored the articulatory mechanisms of HBB in the vocal tract mid-sagittal plane. Proctor et al. (2013) analyzed the articulatory mechanisms of 17 HBB drum sounds belonging to the repertoire of a professional beatboxer. They found that they were similar to those exploited in speech, such that the authors were able to annotate each sound using the International Phonetic Alphabet (IPA) which was originally devised to represent the sounds of spoken languages. Shared characteristics in terms of manner (stops, fricatives, and affricates) and place of articulation (oro- and laryngopharyngeal regions) were described in an acoustic study (Yeshoda & Raveendran, 2021). However, further data (Blaylock et al., 2017) showed that beatboxers employ an extremely wide variety of articulatory mechanisms, in terms of both place and manner of articulation, as well as airstream mechanisms, often non-attested in speech but some of which were recently mentioned in vocal imitations of non-speech sounds (Friberg et al., 2018; Helgason, 2014). In addition, the higher the level of expertise, the better the control of articulatory and airstream mechanisms (Patil et al., 2017). All three studies (Blaylock et al., 2017; Patil et al., 2017; Proctor et al., 2013) described the use of ejective productions of several plosive sounds, wherein the closed glottis acts like an upward moving piston to compress the air trapped between the glottal closure and a supraglottal closure (Ladefoged & Maddieson, 1996) to produce a more intense sound upon supraglottal closure release than a pulmonic plosive would produce. More recently, Dehais-Underdown et al. (2021) conducted an aerodynamic, acoustic, and laryngoscopic investigation of 9 drum imitations produced by 5 beatboxers and were able to attest the use of all the 6 physiologically possible initiation mechanisms. Further, they observed a combination of glottalic and pulmonic airstreams for the same sound (an ingressive affricate).

It is clear that scientific evidence has proven the use of a wider range of production mechanisms in HBB than speech and refined pneumo-phono-articulatory coordination.

However, thus far we only have scratched the surface and our understanding of the production mechanisms of HBB remain superficial. Among other aspects, the similarities and differences between HBB and speech production are not well understood.

Part II

Methodological framework

Chapter 3

Materials and Methods

Contents

3.1 Tools

The first studies on HBB production (mostly undergraduate work) have exploited only audio signals for acoustic investigations on HBB sounds (for a more in-depth review, see Pillot-Loiseau et al., 2021; Pillot-Loiseau et al., 2020). More recently, it has been more common to associate at least two recording techniques, with some exception (Yeshoda & Raveendran, 2021). Acoustics has been associated with aerodynamic measurements and endoscopy (Dehais-Underdown et al., 2021). However, the most exploited by far is imagery: multiple studies use endoscopy (De Torcy et al., 2014; Dehais-Underdown et al., 2021; Sapthavee et al., 2014), or real-time magnetic resonance imaging or rtMRI (Blaylock et al., 2017; Patil et al., 2017; Proctor et al., 2013). Endoscopy has the great advantage of directly visualizing the laryngeal and pharyngeal structures, however the visualization is in 2 dimensions, and sometimes the retraction of the tongue root can obscure the view. Further, it is a medical procedure and must be performed by a trained physician in a medical environment. MRI is the only imaging technique that displays the integrity of the VT, although only on one plane (in HBB studies, the midsagittal plane). Some

inconveniences with MRI are the supine position of the beatboxer, a fairly low temporal resolution and the noise of the machine, that, other than being uncomfortable, disturbs the audio recordings. Fortunately, the two latter issues can be dealt with in post-processing (Patil et al., 2017).

However, more recording techniques are available, that provide valuable data for the characterization of the production mechanisms of HBB sounds. These techniques are well known and have been employed in many other other fields, such as speech and singing, but have never been used to investigate HBB. We deemed it important to have a more comprehensive picture of multiple synchronous physiological information to better understand the mechanisms of production of HBB sounds. Hence, for our investigations we chose to use multiple synchronized techniques to gather multiple physiological data and complement the available knowledge on HBB production mechanisms. We combined electromagnetic articulography, respiratory inductive plethysmography, surface electromyography, electroglottography, acoustic and video recordings.

Electromagnetic articulography (EMA) is widely used in speech research (Barbier et al., 2020; Brunner et al., 2010; Rebernik et al., 2021; Savariaux et al., 2017; Tiede et al., 2019a), providing valuable and quantifiable information on actual flesh points. It measures the position (a triplet of coordinates x, y, z) over time at very high temporal resolution of sensor coils attached to the lips, tongue and other parts of the mouth while they move in an electromagnetic field. We turned to EMA to gather direct information on the movements and speeds of the articulators to disclose further details on HBB production mechanisms. However, coils may become detached after intense or prolonged recording sessions. Given the nature of HBB production, we were not sure whether EMA coils would withstand a recording session of HBB and if beatboxing would be achievable despite the presence of the coils on the articulators, especially the tongue.

Respiratory inductive plethysmography (RIP) is used to evaluate pulmonary ventilation by measuring thoracic and abdominal cross sectional area changes. Two pairs of sinusoid wire coils are placed around the thorax and the abdomen. The variations of the cross sectional area of these two compartments during breathing alters the self-inductance of the coils. The major downside of this technique is that tidal volume is indirectly estimated from the variations of the cross sectional area, in other words, RIP measure cumulative variables (Traser, 2017). However, we opted for RIP because it would provide reliable information on respiratory behavior synchronous with all other measurements without the need of further captors around mouth. Additionally, the coils being secured in a fitting vest, the beatboxers had a relative freedom of movements.

In surface Electromyography (sEMG), two electrodes measure the difference in action potential between two neighbouring points of the same set of muscle fibres, giving an indication of muscle activation. We were particularly interested in this data because no

information is available on timing and intensity of muscular activity in HBB.

Electroglottography (EGG) measures the degree of contact between the vocal folds. The EGG signal is composed of both a high-frequency component, which reflects vibration of the vocal folds (voicing) and a low-frequency component corresponding to slow vertical motions of the larynx (e.g., during swallowing). For a recent review on EGG use in research, see Herbst (2020).

3.2 Corpora

The present work is grounded on data drawn from three corpuses collected over a span of three years with six beatboxers of different level of expertise. In the following sections, the three corpuses are presented.

3.2.1 Pilot corpus

This pilot corpus C1 was constituted in 2016 at GIPSA-lab as a first attempt at recording EMA data on a beatboxer. For more details, see Paroni 2016, Paroni 2018.

The protocol (3.2.1.2) was designed to investigate similarities and differences in articulatory and breathing behavior as well as acoustic outcome of the occlusive consonants [p, t, k] and the corresponding HBB sounds (P, t, K).

3.2.1.1 Participant

The participant (PS) is a 28 year-old left-handed male, native speaker of French. He has been practicing HBB for 9 years at an amateur level. He occasionally performs in concerts, however he has never participated in official HBB competitions. Because of having often experienced vocal fatigue and discomfort after practicing, PS learned the diaphragmatic breathing technique (Leanderson & Sundberg, 1988; Leanderson et al., 1984). He reports benefiting from this breathing technique in his HBB practice, and that he no longer suffers from vocal fatigue.

3.2.1.2 Protocol

The protocol started with an interview of the beatboxer prior to the recording session. His experience in HBB and his vocal habits were collected. The experimental details were

presented to him. The HBB effects of interest for the study – five categories of drum sounds (kick, hi-hat, snare, rimshot and cymbals) and their variants – were discussed with him. He stated that he could produce more than one sound for each category: a humming variant and a non-humming one (that he called *power*), an inhaled variant and an exhaled one. All the humming sounds were produced without superimposed melody or voicing (see audio files available as supplementary material). [1]

The protocol was organized in three sections:

I **Drum sounds**: 12 sounds belonging to five categories of vocal drum set sounds were recorded:

kick humming and power variants

hi-hat humming and power variants, open and closed for the power variant

snare humming and power variant, exhaled and inhaled for the power variant

rimshot humming and power variants

cymbal exhaled and inhaled

Each sound was repeated at least 15 times, while following the tempo provided by a metronome set at 80 beats per minute (bpm), and varying loudness when possible.

II **Speech syllables**: three speech syllables were recorded to be compared to the three corresponding HBB sounds:

pu corresponding to the HBB kick sound

ti corresponding to the HBB closed hi-hat sound

ka corresponding to the HBB rimshot sound

The syllables were produced in three speech modes (normal speech, shout and whisper) and in their HBB counterpart. They were repeated at least 30 times, varying vocal intensity and following the tempo provided by a metronome set at 80 bpm;

III **Sentences**: Three sentences:

des petits cookies des gros cookies [depətiku'kidegʁoku'ki], English translation: little cookies, big cookies

pâtes au pesto ['patopes'to], English translation: pesto pasta

boots and cats French pronunciation ['butsɛn'kats]

[1] Supplementary materials are available online at doi.org/10.5281/zenodo.4264746

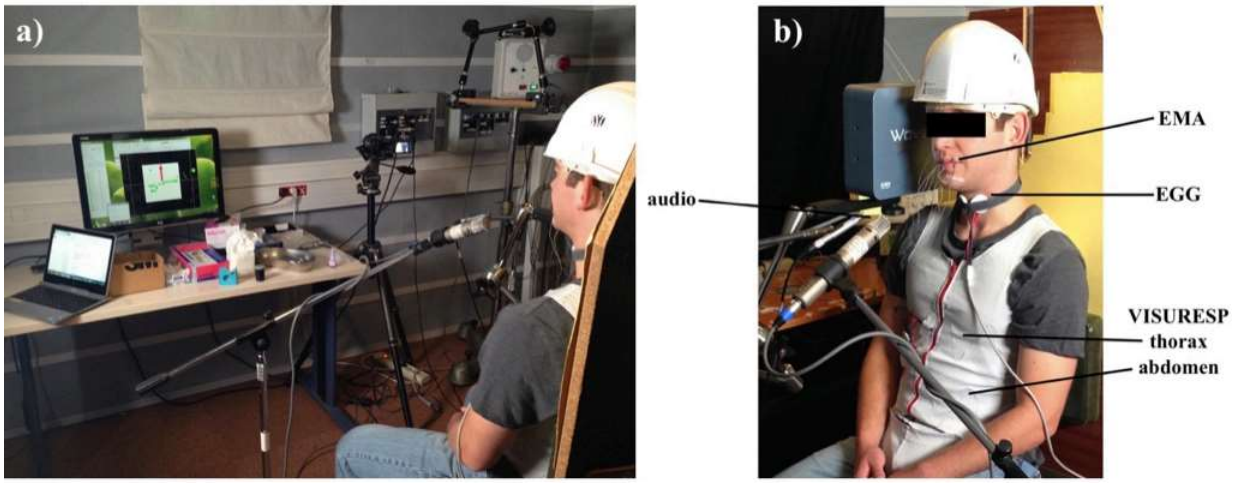


Figure 3.1: a) Experimental setting and b) apparatus.

were produced on a continuum going from normal speech to HBB, and from HBB back to normal speech.

The choice of the syllables and phrases was determined by the fact that they are meaningful to beatboxers, in that they often constitute the basis for HBB learning.

3.2.1.3 Experimental setting and apparatus

The recordings took place in the semi-anechoic room of GIPSA-lab in Grenoble, a place of biomedical research authorized by the ARS Auvergne-Rhône-Alpes.

After being interviewed and signing an informed consent form, the subject was placed in the recording room (Fig.3.1a), wearing a waistcoat for respiratory inductance plethysmography (VISURESP system, RBI, France), and sitting on an adapted chair that assured the stabilization of the head inside of the magnetic field of an electromagnetic articulograph (EMA WAVE, NDI, Canada) (Fig.3.1b). To collect the articulatory data, 12 coils were positioned as follows (Fig.3.2):

- 3 coils were placed midsagittaly on the tongue: 1 coil about 1 cm from apex (TIP), 1 coil on the blade about 3 cm from apex (MID), 1 coil on the dorsum about 5 cm from apex (DORS);
- 1 coil on the medial lower incisors (JAW);
- 2 coils on the upper lip (mid, UML and left, ULL), 2 coils on the lower lip (mid, LML and left, LLL);

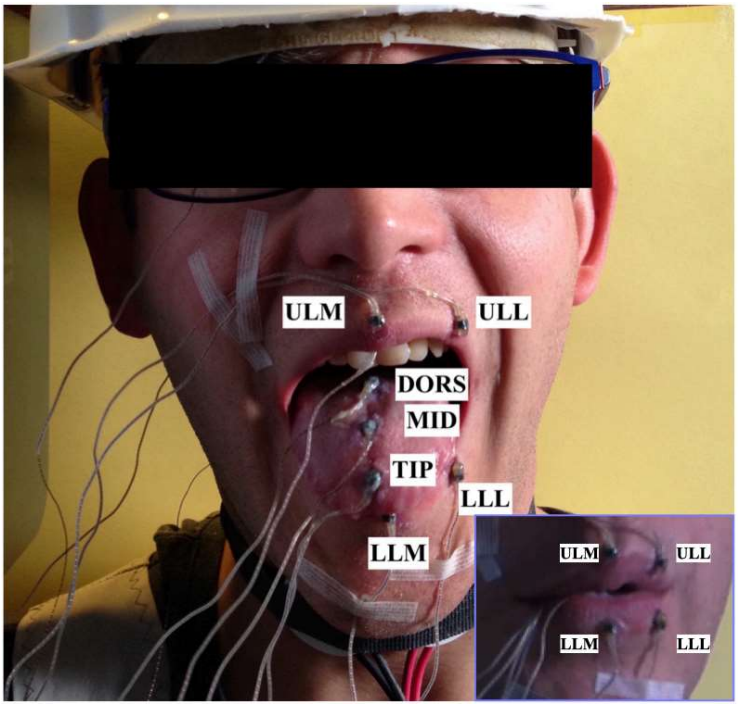


Figure 3.2: Coil placement on the tongue and lips.

- 1 reference coil on the upper incisors;
- 2 reference coils on the mastoid processes behind both right and left ears;
- 1 reference coil on the nasion.

The EMA signal was sampled at 400 Hz. After recording, EMA data were post-processed in two steps (for more details on the method, see Tiede et al., 2019b). As a first step, movement of the head was corrected with Matlab software using the 4 reference coils glued on the nose, upper incisor, and behind both ears. As a second step, a rotation and a translation were applied to reference the data in the coordinate system of the beatboxer. Figure 3.3 shows an example of EMA data after post-processing.

Two pairs of electrodes (Glottal Enterprise EG2 dual-channel electroglottograph, Rothenberg, 1992) were positioned on the neck of the subject in the larynx region (Fig. 3.1 b) for measuring vocal-fold contact and detecting laryngeal movements. An AKG microphone and a 1/2" prepolarized free-field microphone (B&K 4189) connected to a microphone preamplifier (B&K 2669C) and NEXUS conditioning amplifier (B&K 2690) were placed at a distance of approximately 20 cm from the subject's mouth in order to capture the audio signal and derive intensity level after calibration. Both electroglottographic (EGG) and audio signals were sent to a BIOPAC unit (MP150) and sampled at 40 kHz. The respiratory inductance plethysmographic (RIP) signals were recorded on two devices: a computer

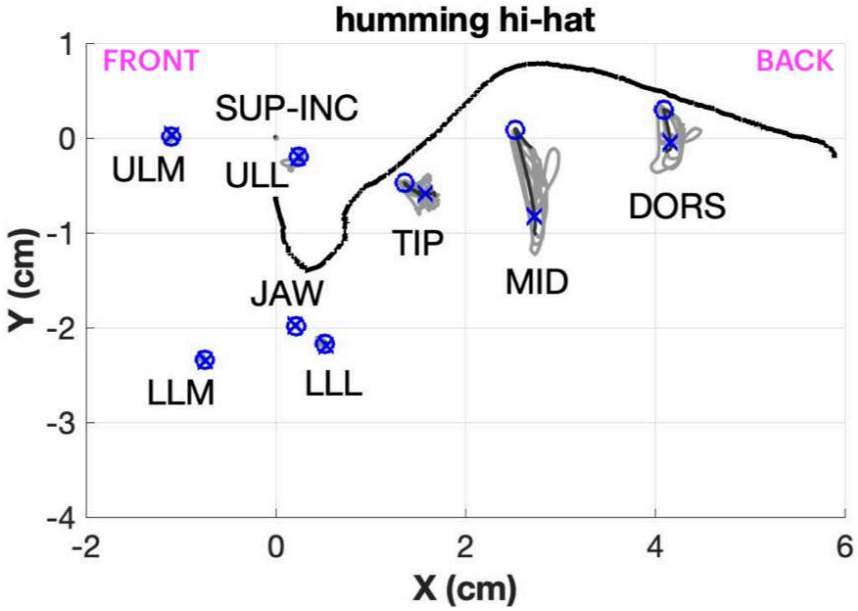


Figure 3.3: Sagittal (XY) view of hard palate contour (black solid line), coil trajectories and corresponding labeling. Front of the oral cavity is on the left. Cross: acoustic burst; circle: extinction of acoustic activity.

dedicated to VISURESP system (at 40 Hz sampling frequency) and the BIOPAC unit so as to be synchronized with audio and EGG signals (at 40 kHz sampling frequency).

A camera was facing the subject for the video recordings at 25 fps. During recording, an acoustic trigger signal (20 ms length square wave) was manually launched by an external electronic device and captured by each system prior to and after each task, so as to allow data synchronization in post-processing.

At the end of the recording session, a coil manually traced the midsagittal plane from the back of the palate to the front of the upper incisors to obtain the palatal contour.

3.2.2 Multi-subject corpus

After assessing the feasibility of EMA technique on HBB production with the constitution of C1, a multi-subject corpus (C2) was recorded. Five beatboxers were recorded (3.2.2.1). Multiple techniques were used synchronously to obtain physiological data (3.2.2.3). First EMA, then sEMG, were combined with RIP, EGG, audio, and video recordings. The protocol (3.2.2.2) was designed to investigate the research questions exposed in the introduction to this work. The experimental protocol and procedure were validated by the Ethics Committee for Research Grenoble Alpes (CER Grenoble Alpes-Avis-2019-06-11-4).

3.2.2.1 Participants

Five male French speaking beatboxers (S01-S05) were recorded, four professionals and one amateur[2], aged 20 to 38 years. S01 is a 35 year-old right-handed male. His first language is French. He is a full-time professional artist, who has been doing HBB for 20 years, winning international competitions and practicing HBB on a daily basis. S02 is a 21 year-old French native speaker, he is right-handed. He is the least experienced beatboxer of the corpus, with 5 years of practice at an amateur level, but having participated in a French national competition nonetheless. He practices HBB every day. S03 is 38 years of age and right-handed, native speaker of French. He is a professional artist, and has participated and won international competitions in the past, but is still active in the French HBB community and still exercises 1-3 times a week. S04 is a 31 year-old professional artist, native speaker of French, and left-handed. He regularly participates in national HBB competitions, although never having won any title so far. He practices his HBB skills every day. S05 is a 20 year-old right-handed native speaker of French. Despite being the youngest beatboxer in our recordings, he regularly participates in official competitions and already is a national HBB champion. He exercises his HBB skills on a daily basis. All beatboxers state that they are self-taught.

3.2.2.2 Protocol

The protocol started with an interview of the beatboxer prior to the recording session. His personal data, experience in HBB, and vocal habits were collected (see Table 3.1). The experimental details were presented to him. The HBB sounds of interest for the study – belonging to five categories of drum sounds (kick, hi-hat, snare, rimshot and cymbals) – were discussed with him. If he stated that he was not familiar with the sound, an audio example was played. Before the recordings, the beatboxer was presented with a tempo at 80 bpm provided by a metronome and asked to use this tempo throughout the protocol.

The protocol itself was organized in two sections:

I **Comparison HBB and speech**: regular kick, hi-hat and rimshot were compared to the syllables [pu, ti, ka], and humming kick, hi-hat and rimshot to the syllables [bu, di, ga] in different phonetic contexts:

/pu/, /ti/, /ka/, /bu/, /di/, /ga/ in isolation

P, t, K regular and humming, in isolation

/putikati/, /budigadi/ the three syllables were combined in a sequence

[2]This professional-amateur classification reflects the level of the beatboxer at the time of the recordings.

PtKt regular and humming, the three sounds were combined in a *beat*

le poulet le ticket le cadeau in sequence ([ləpuˈlɛlətiˈkɛləkaˈdo], English translation: the chicken, the ticket, the gift)

le boulet le dico le gâteau in sequence ([ləbuˈlɛlədiˈkoləgaˈto], English translation: the cannonball, the dico – the name of a dictionary –, the cake)

le P le t le K regular and humming, in sequence Each item was repeated 12 times. Each sequence of syllables and sounds in isolation was introduced by [sasɛlə] (English translation: this is the).

II **Drum sounds**: 27 HBB sounds were recorded, among 5 categories of vocal drum sounds:

kick Classic kick, Humming classic kick, Reverse classic kick, Inward classic kick, Classic kick roll, Dry kick

snare Classic snare drum, Humming classic snare drum, Inward K snare, 808 snare, 808 snare roll, Esh snare, Dry snare, Outward K snare, Percussive K snare

rimshot Inward rimshot

hi-hat Open hi-hat, Humming open hi-hat, Closed hi-hat, Humming closed hi-hat, Reverse open hi-hat, Fast hi-hat, Reverse fast hi-hat, Reverse snare hut

cymbal Brushed cymbal, Splash cymbal, Crash cymbal

Each sound was repeated 12 times, the sequence being introduced by [sasɛlə] (English translation: this is the).

The first section of the corpus was recorded twice, one to collect EMA data, the other to collect sEMG data, the placement of the captors not allowing simultaneous use of the two techniques.

3.2.2.3 Experimental setting and apparatus

Similarly to the pilot corpus (see 3.2.1.3), the recordings took place in the semi-anechoic room of GIPSA-lab in Grenoble.

The five participants (S01-S05) were recorded on five different sessions, each taking place on a different day. On the day S01 underwent his session, the EMA system was out of order, and therefore no EMA data were collected. The procedure was similar to that described in the case of C1 (see 3.2.1.3): after signing an informed consent form, the participant was placed in the recording room (Fig. 3.4), wearing a fitted T-shirt for RIP (ETISENSE, France), and sitting on a chair. He was asked not to lean on the backrest

in order to avoid disruption of RIP data acquisition. The RIP signals were recorded at 200 Hz on a computer dedicated to ETISENSE system.

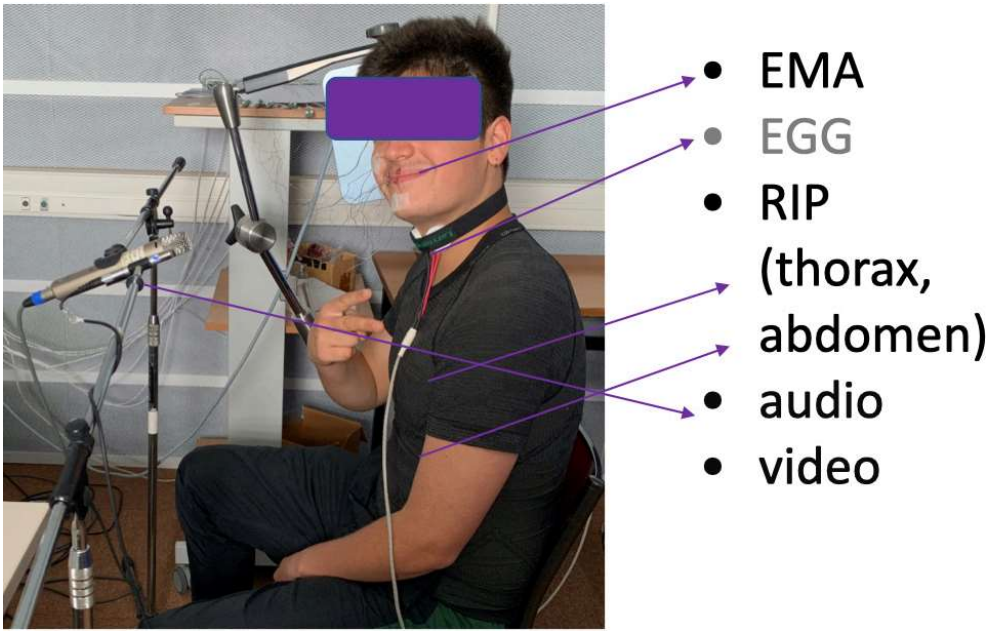


Figure 3.4: Experimental setting and apparatus.

Two pairs of electrodes (Glottal Enterprise EG2 dual-channel electroglottograph, Rothenberg, 1992) were positioned on the neck of the beatboxer in the larynx region (Fig. 3.4) for measuring vocal-fold contact and detecting laryngeal movements. An AKG microphone and a 1/2" prepolarized free-field microphone (B&K 4189) connected to a microphone preamplifier (B&K 2669C) and NEXUS conditioning amplifier (B&K 2690) were placed at a distance of approximately 20 cm from the subject's mouth in order to capture the audio signal and derive intensity level after calibration. Both EGG and audio signals were sent to a BIOPAC unit (MP150) and sampled at 20 kHz.

A camera was facing the subject for the video recordings at 25 fps.

The beatboxer's head was placed in the magnetic field of an articulograph (EMA WAVE, NDI, Canada). He was asked to be mindful of the position of his head with respect to the the articulograph, in order not to exit the magnetic field. To collect the articulatory data, 10 coils were placed as follows (Fig. 3.5)

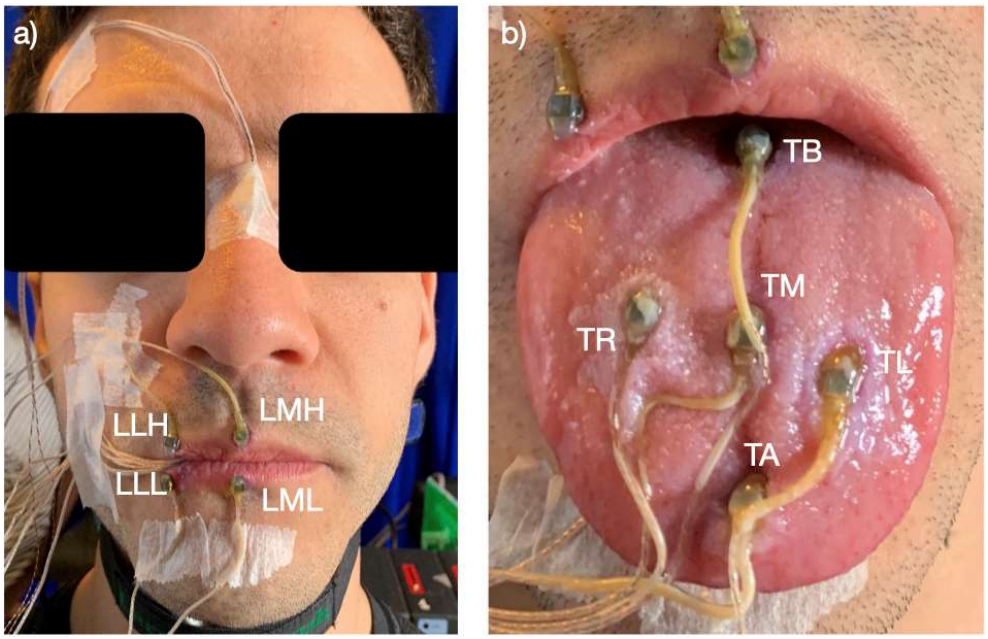


Figure 3.5: Coil placement a) on the lips and b) on the tongue.

- 3 coils were placed midsagittaly on the tongue:
  - 1 coil about 1 cm from apex (TA),
  - 1 coil on the blade about 3 cm from apex (TM),
  - 1 coil on the dorsum about 5 cm from apex (TB);
- 2 coils were placed laterally on the tongue:
  - one on the right (TR)
  - the other on the left (TR) of TM;
- 4 coils on the lips
  - 1 coil in the mid portion of the upper lip (LMH),
  - 1 coil in the mid portion of the lower lip (LML),
  - 1 coil in the right (RLH) or left (LLH) portion portion of the upper lip,
  - 1 coil in the right (RLL) or left (LLL) portion portion of the lower lip[3].

[3]The coils were placed on the right or left portion of the lips based on two criteria: the beatboxer was asked if he released the occlusion of the kick centrally or laterally. If the answer was laterally, the coil was placed on the side of the release. If the answer was centrally, the coil was placed on the side matching the handedness of the beatboxer.

- 1 coil on the medial lower incisors (JAW);
- 1 reference coil on the nasion.

The EMA signal was sampled at 100 Hz for S04 and at 200 Hz for S01, S02, S03, and S05.

After going through the whole protocol, a coil was used to manually trace the mid-sagittal and the 3D plane from the back of the palate to the front of the upper incisors to obtain the palatal contour. The bite plane was also acquired.

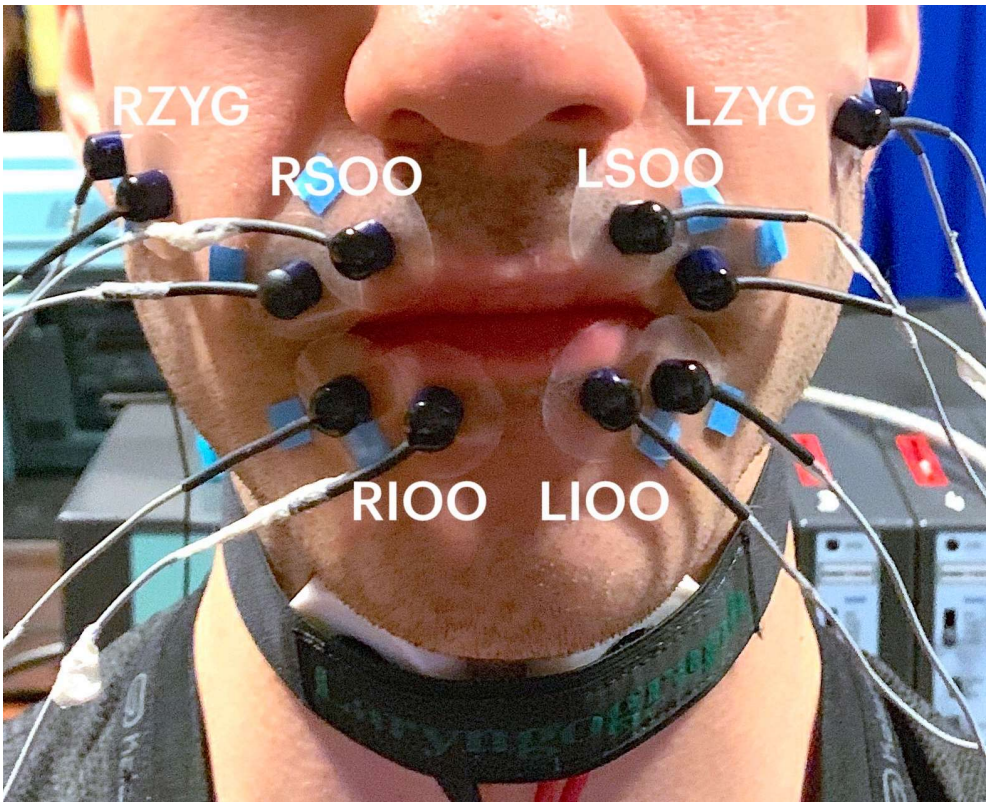


Figure 3.6: Electrodes placement.

The EMA coils were then removed and 6 pairs of sEMG electrodes and 1 grounding electrode were placed as such (Fig.3.6):

- 1 pair of electrodes on the right superior orbicularis oris (RSOO),
- 1 pair on the right inferior orbicularis oris (RIOO),
- 1 pair on the left superior orbicularis oris (LSOO),

- 1 pair on the left inferior orbicularis oris (LIOO),
- 1 pair on the right zygomatic major (RZYG),
- 1 pair on the left zygomatic major (LZYG),
- 1 grounding electrode on the right mastoid bone.

The electrode placement was determined by palpation by a trained speech and language therapist (the author) while the beatboxer was asked to press his lips together (orbicularis oris muscle) or to retract the angles of his mouth (zygomatic muscles). The sEMG signal was sent to a BIOPAC unit (MP150) together with EGG and audio signals, and was sampled at 20 kHz.

During recording, an acoustic trigger signal (50 ms, 100 ms, 150 ms, 200 ms, 250 ms length square waves in sequence) was launched by the ETISENSE system and captured by each system prior to and after each task, so as to allow data synchronization in post-processing.

3.2.3 Endoscopy corpus

S02 and S03 participated in the constitution of a corpus of endoscopic and audio data. The nasofibroscopy was performed by a board certified otolaryngologyst at Centre Hôspitalier Universitaire in Grenoble in 2018. The protocol was designed to compare speech consonants [p, b, t, k] to HBB kick (P), hi-hat (T), and rimshot (K). The sentences:

boots and cats French pronunciation [ˈbutsɛnˈkats]

pâtes au pesto [patoˈpesto], English translation: pesto pasta

pose ta capuche, t'as qu'à poser ta capuche [ˈpoztakaˈpyʃtakapoˈzetakaˈpyʃ][4], English translation: put your hood down, just put your hood down

and their beatboxed equivalents:

P T K E-H

P T PS E-H

[4]In actual facts, the third stress was rather shifted on the seventh syllable: [ˈpoztakaˈpyʃtakaˈpozetakaˈpyʃ].

PS T K PS T K PS T K PS

were recorded at different enunciation speeds.

For more details, see Fabre, 2018.

3.2.4 Corpora at a glance

For the reader's convenience, the main structure of the three corpora are schematized here. Table 3.1 gives an overview on the participants, Table 3.2 on the recording techniques, and Talbe/Figure summarizes the protocols.

Table 3.1: Global overview of the beatboxers recorded in the three corpuses. The beatboxer is regarded as a professional if he earns his living from his practice. The competition level is based on participation in official competitions: (1) never or <2; (2) ≥ 2, no wins; (3) ≥ 2, with wins.

Subject	Corpus	Sex and Age	Laterality	Expertise	Competition	First language
PS	C1	M26	L	amateur	1	French
S01	C2	M35	R	pro	3	French
S02	C2, C3	M21	R	amateur	1	French
S03	C2, C3	M38	R	pro	3	French
S04	C2	M31	L	pro	2	French
S05	C2	M20	R	amateur	3	French

Table 3.2: Global overview of the three corpuses.

Corpus	Subjects	Type	Data
C1	1	Multi-physiology (pilot)	EMA, RIP, EGG, audio, video
C2	5	Multi-physiology	EMA, RIP, sEMG, EGG, audio, video
C3	2	Endoscopy	endoscopic, audio

3.3 Analyses

Post-processing of the data consisted of firstly a rototranslation of the EMA data with respect to the bite plane and then the synchronization of all the recorded data.

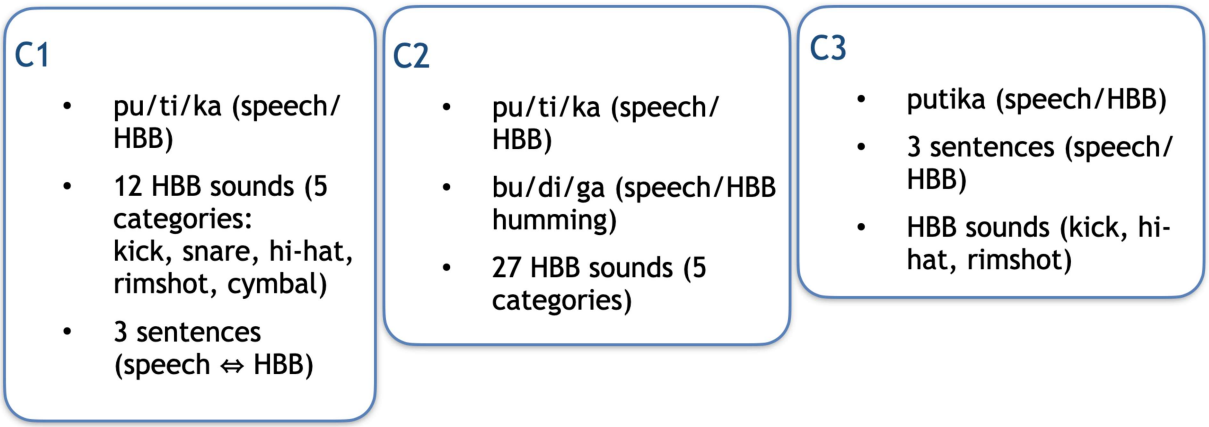


Figure 3.7: Brief overview of the items recorded in each corpus.

3.3.1 Segmentation and annotation

All the analyses were based on timestamps detected on the audio and EGG signals recorded by the BIPAC unit. The signals were segmented and annotated using the commercial software Praat (Boersma, 2006). A TextGrid was created for each item. The first tier was used to indicate the mode of production (C1: speech, transition, HBB; C2: speech, HBB). The criteria to place the boundaries were different depending on the item. When a carrier sentence was present (Fig. 3.8 and 3.9 C2 [sasɛlə]), the left boundary was placed in correspondence with the last visible oscillation of the vowel [ə] and the right boundary in correspondence with the last oscillation of the last vowel for speech sequences and at the sound extinction for HBB sequences. The second tier was used to annotate the item and the third tier was used for phonetic annotation, following the criteria: consonants were segmented from acoustic burst to voice onset time (VOT), vowels from the first to the last oscillation detected on the EGG signal, HBB sounds from burst to sound extinction. The presence or absence of voicing in HBB was annotated on a separate tier for HBB production. A last point tier was created for ease of analysis where only the times of burst were marked.

The analyses subsequently performed are presented with more detail in each chapter.

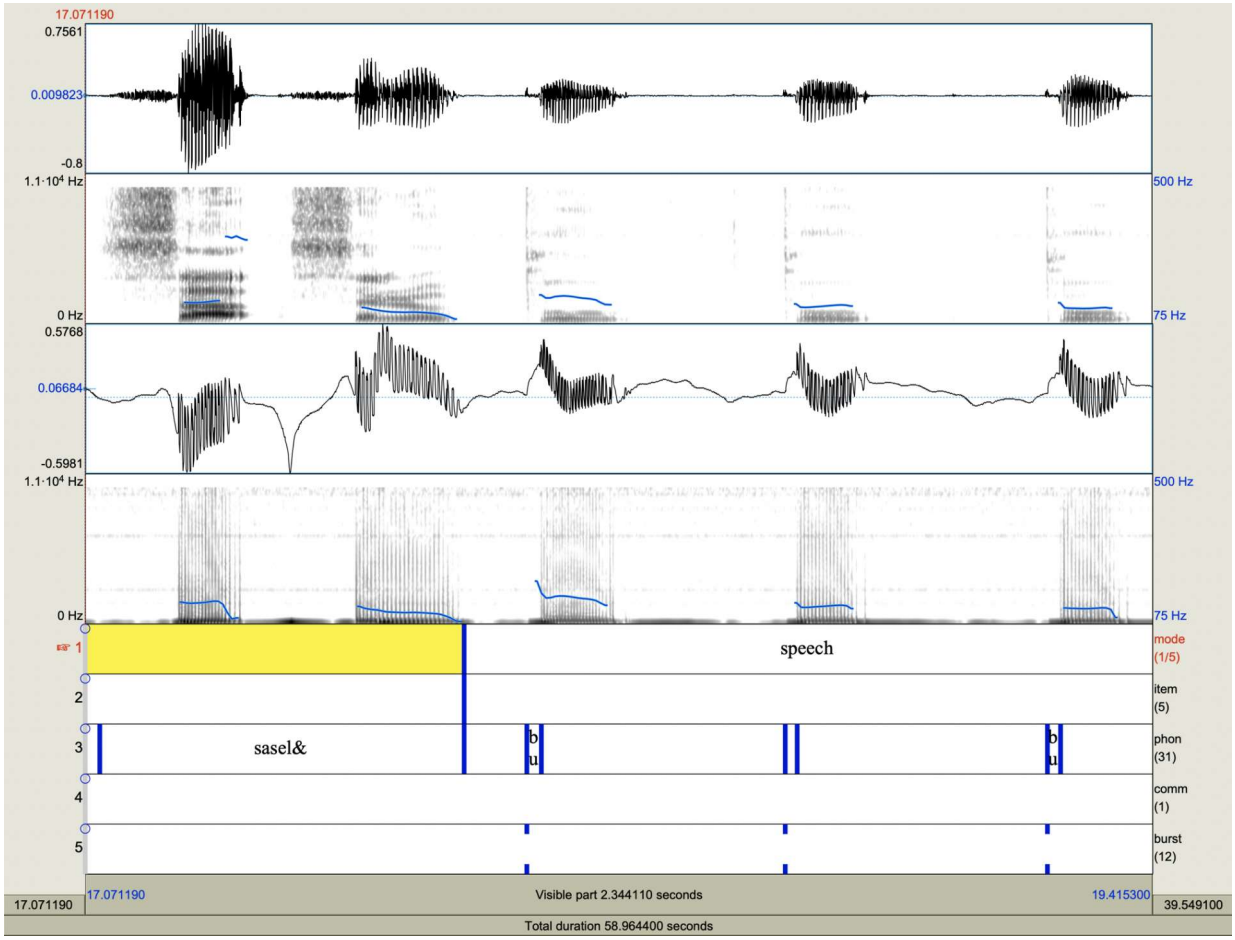


Figure 3.8: Example of segmentation and annotation of the acoustic and EGG signal relative to speech produced by S04 during the task of repetition of the syllable /pu/.

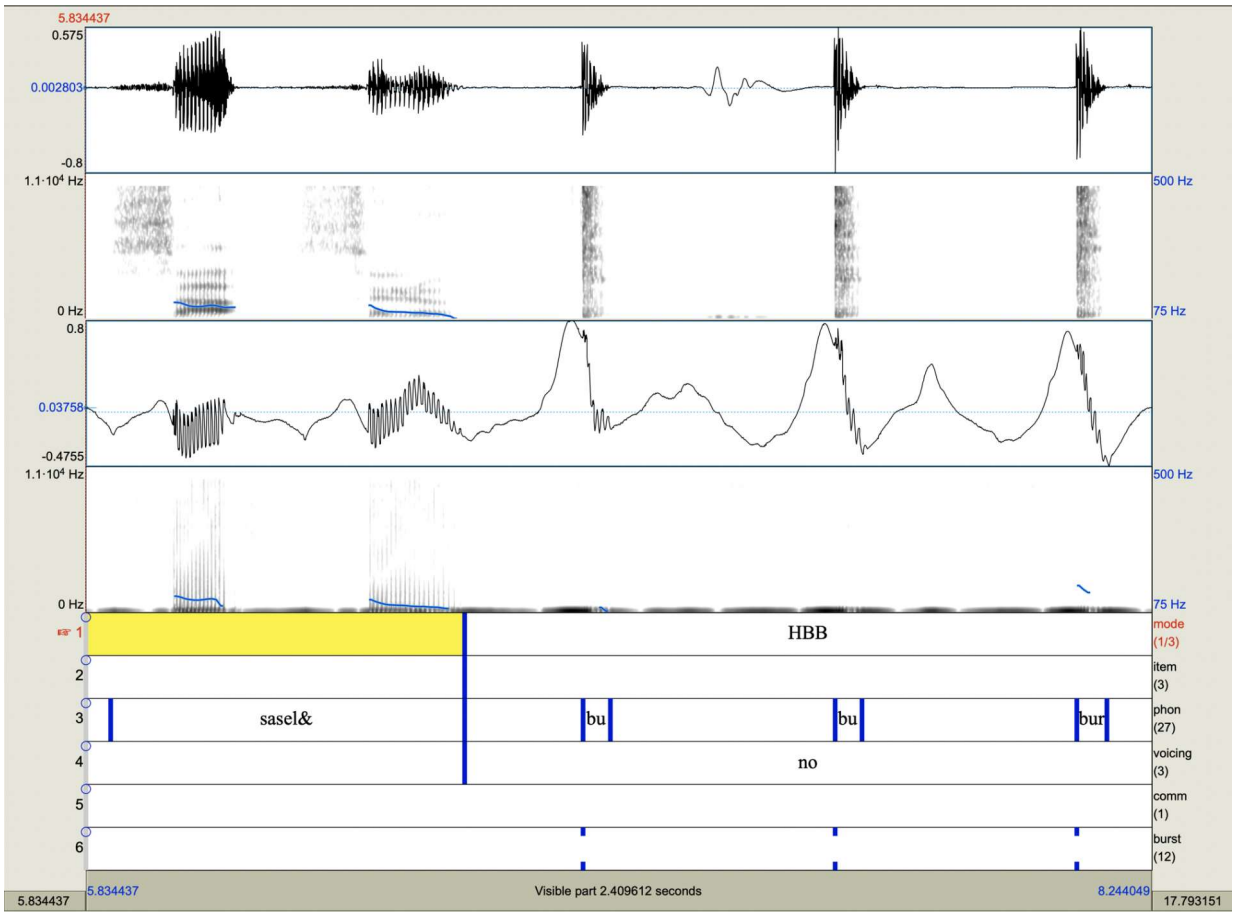


Figure 3.9: Example of segmentation and annotation of the acoustic and EGG signal relative to HBB produced by S04 during the task of repetition of the kick.

Part III

Experimental part

Chapter 4

Beatboxing, the basics: Drum set sounds

Contents

HBB basics in a nutshell is the imitation of drum sounds. Often the first sounds to be learned, they constantly have to be worked and be on point. Clarity of sound, i.e., the acoustic outcome is a paramount parameter for performance evaluation at every stage of competition, including world championships. HBB is not only originality and extravagance of sounds, but primarily mastering of the basics. From an artistic standpoint, this comes down to a very accurate imitation of drum sounds. But from a scientific standpoint,

what are the characteristics of these sounds? What production mechanisms are used to imitate drum sounds? What acoustic characteristics do HBB sounds have? Sounds deemed different from a musical standpoint (e.g., a K snare and an inward K snare) are different also by a scientific standpoint? How do we, as scholars, tackle the issue of annotating non-linguistic sounds? Can we use phonetic annotation systems designed for speech sounds?[1]

4.1 Drum set sounds

The few available studies so far that investigate HBB production mechanisms have employed techniques such as endoscopy (De Torcy et al., 2014; Dehais Underdown et al., 2019; Sapthavee et al., 2014) and rtMRI (Blaylock et al., 2017; Patil et al., 2017; Proctor et al., 2013). While providing valuable information on the general behavior of the articulators, neither technique allows the study of the dynamics of a given flesh point on an articulator. Both techniques are also limited by a relatively low sampling frequency. More precise and quantitative evaluation of the articulatory dynamics could be performed using EMA, a widely-used technique in speech research to measure the position and movement over time of selected points on articulators (Barbier et al., 2020; Brunner et al., 2010; Savariaux et al., 2017; Tiede et al., 2019a).

This study is part of an ongoing effort to understand the production of HBB drum sounds by exploring lingual and labial articulatory dynamics in relation to acoustic characteristics and ventilatory behavior on a beatboxer producing five categories of drum sounds belonging to his repertoire (kick, snare, hi-hat, rimshot, cymbal). We rely on the EMA technique to explore the kinematics of tongue and lip-flesh points. Acoustic and articulatory characterizations provided in Sections 4.3.1 and 4.3.2 lead to a phonetic description of the ways these vocal drum sounds are produced. Section 4.4 discusses the specifics of HBB sound production.

4.2 Methods

The results presented in this chapter are drawn from C1.I.

Table 4.1 summarizes the participant, the items, and the techniques relative to the data

[1]This chapter is part of a published paper reformatted to meet the needs of this dissertation. Source: Paroni, A., Henrich Bernardoni, N., Savariaux, C., Lœvenbruck, H., Calabrese, P., Pellegrini, T., Mouysset, S., and 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, pp. 191–206. Available at: https://asa-scitation-org.sid2nomade-1.grenet.fr/doi/full/10.1121/10.0002921

used to conduct preliminary observations (Sec. 5.3). For more details, see section 3.2.1.

Table 4.1: Visual summary of participant, items, and techniques employed.

Participants	Items	Variant	Techniques
PS	kick	power/humming	EMA,
	hi-hat	power/humming	RIP,
	snare	power (out/in)/humming	EGG,
	rimshot	power/humming	audio,
	cymbal	out/in	video

Audio files were manually segmented and phonetically annotated using the software Praat (version 6.0.49, Boersma, 2006). The phonetic annotations were carried out inspecting audio, video and EGG data. Audio files were segmented using the following criteria: the left boundary was placed in correspondence with the burst and the right boundary in correspondence with the last visible oscillation on the waveform. Boundaries subsequently provided timestamps for the meaningful quantities investigated. The phonetic annotations were performed by the first author who is a speech therapist and has also received a training as a linguist. The alphabet used was the Worldbet Alphabet (Hieronymus, 1993), which is the translation of IPA into symbols compatible with automatic data processing.

A clustering technique[2] was used to test whether HBB sounds are distinguishable on the basis of the acoustic signal. Spectral clustering on 12 t-SNE-whitened Mel Frequency Cepstral Coefficients (MFCCs) was applied. t-SNE (Maaten & Hinton, 2008), which stands for t–Distributed Stochastic Neighbor Embedding, is a recent and efficient non–linear projection technique (SC, Von Luxburg, 2007). The first coefficient (C0) was removed, as it measures signal loudness that is not relevant to characterize the frequency content of interest. The MFCCs were extracted every 6.25 ms on 25 ms duration frames, with 50 Hz and 8000 Hz as minimum and maximum extreme frequency values to compute the Mel bands.

EMA data were processed using the commercial software package MATLAB "MathWorks: Bioinformatics Toolbox: User's Guide (R2018b)," 2018. The spatial trajectories of the 8 coils positioned on the tongue, jaw and lips were computed. A visual inspection of the trajectories was carried out to characterize the articulation of each HBB sound. Corrections to the phonetic annotation were introduced when needed.

The EGG signal was visually inspected to detect vocal fold vibration phases.

[2]This analysis was performed by the sixth and seventh author.

RIP data were calibrated following the method used and described by Eberhard et al., 2001 and Calabrese et al., 2007. The thoracic and abdominal signals measured with RIP were simultaneously recorded, together with the airflow signal measured by a flowmeter (Fleishhead no.1, Emka Technologies, Paris, France), a differential transducer (163PC01D36, Micro Switch, Honeywell, United States), and a face mask worn by the subject while breathing spontaneously for approximately one minute. Thoracic and abdominal signals recorded with RIP were subsequently linearly combined to obtain the ventilatory volume (VR) signal. The linear coefficients were estimated from the least square method to fit the airflow signal recorded with the flowmeter.

Several parameters were extracted from the annotated data: sound duration and vocal intensity (from acoustic signal), maximum of tangential speed and acceleration (from EMA signals). Three statistical analyses were performed using the R software (R Core Team, 2013). Firstly, a test was run to inspect if a difference in intensity (response variable, in logarithmic scale) exists between variants (humming vs. power) of the same effect (kick, snare, hi-hat, rimshot). Secondly, an analysis was carried out to test what kind of relationship exists between duration (response variable, in logarithmic scale) and intensity in each HBB sound (12 modalities: humming kick, humming snare, humming hi-hat, humming rimshot, power kick, power snare, power inward snare, power closed hi-hat, power open hi-hat, inhaled cymbal, exhaled cymbal). Lastly, an analysis was carried out on the HBB sounds to inspect whether a significant difference exists among the means of the maximum speed of pairings of lingual articulators (TIP, MID, DORS) and of lip articulators (JAW-LLL, LLL-ULL). The considered factors are the coils (8 modalities: TIP, MID, DORSUM, JAW, ULL, UML, LLL, LML) and the HBB sounds (12 modalities: as above) and their interaction. Each analysis was run using the lme function of the nlme package. This function takes into account potential differences in residual variances across HBB vocal drum sounds, or possible correlations among coils in the third analysis. Repetition is considered as a random effect. All the p-values reported in Section are provided by the glht function of the multcomp package (Hothorn et al., 2008) calculated from the corresponding model. For the first and the third model, the estimated differences of the comparisons and their estimated standard errors are provided. For the second analysis, the estimated values of the slopes and their estimated standard errors are provided.

4.3 Results

In this section, prototypical examples are presented. The corresponding audio examples and video files can be found online[3] as supplementary material.

[3]https://zenodo.org/record/4264747#.Yp7ebxNBz0o

4.3.1 Acoustic characterization

341 sound realizations of twelve HBB sounds were analysed. Acoustic characterization performed through spectral clustering achieved a 94% clustering purity value. 19 samples out of the 341 realizations were misclassified. Out of these 19 misclassifications, 12 were annotation errors. For instance, four exhaled cymbal realizations were wrongly annotated as inhaled cymbal. The remaining misclassifications were confusions, among which the most frequent was between humming kick and humming snare. Fig. 4.1 shows the data points after a two-dimension reduction with t-SNE. In this plot, x-axis and y-axis are the output of the t-SNE projection technique and thus, they are arbitrary scales. Each data point is plotted using shape and color according to its sound label (colored version of the figure available online). Pure and meaningful compact clusters can clearly be identified. In general, variants of a same HBB effect are also close together, e.g., the points for power kick and humming kick lie in the same region. Cymbal (in particular the inhaled variant) and hi-hat points are close together, which makes sense, as the two sounds have a similar acoustic signature (see Fig. 4.2 and audio files online).

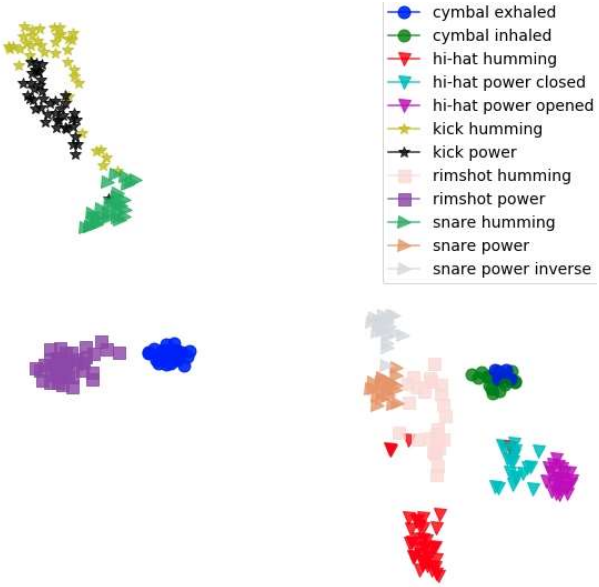


Figure 4.1: Visualization obtained with the t-SNE projection technique. Although the x-axis and y-axis are arbitrary scales, one can see that the different sounds are clearly grouped into distinct clusters (color version available online).

This classification accuracy, *i.e.* the fact that each sound can be correctly assigned to its corresponding cluster via unsupervised methods, demonstrates that each HBB vocal drum sound has its own characteristic acoustic signature. Fig. 4.2 illustrates these signatures with the waveform and spectrogram of a representative token for each HBB sound explored in the present study.

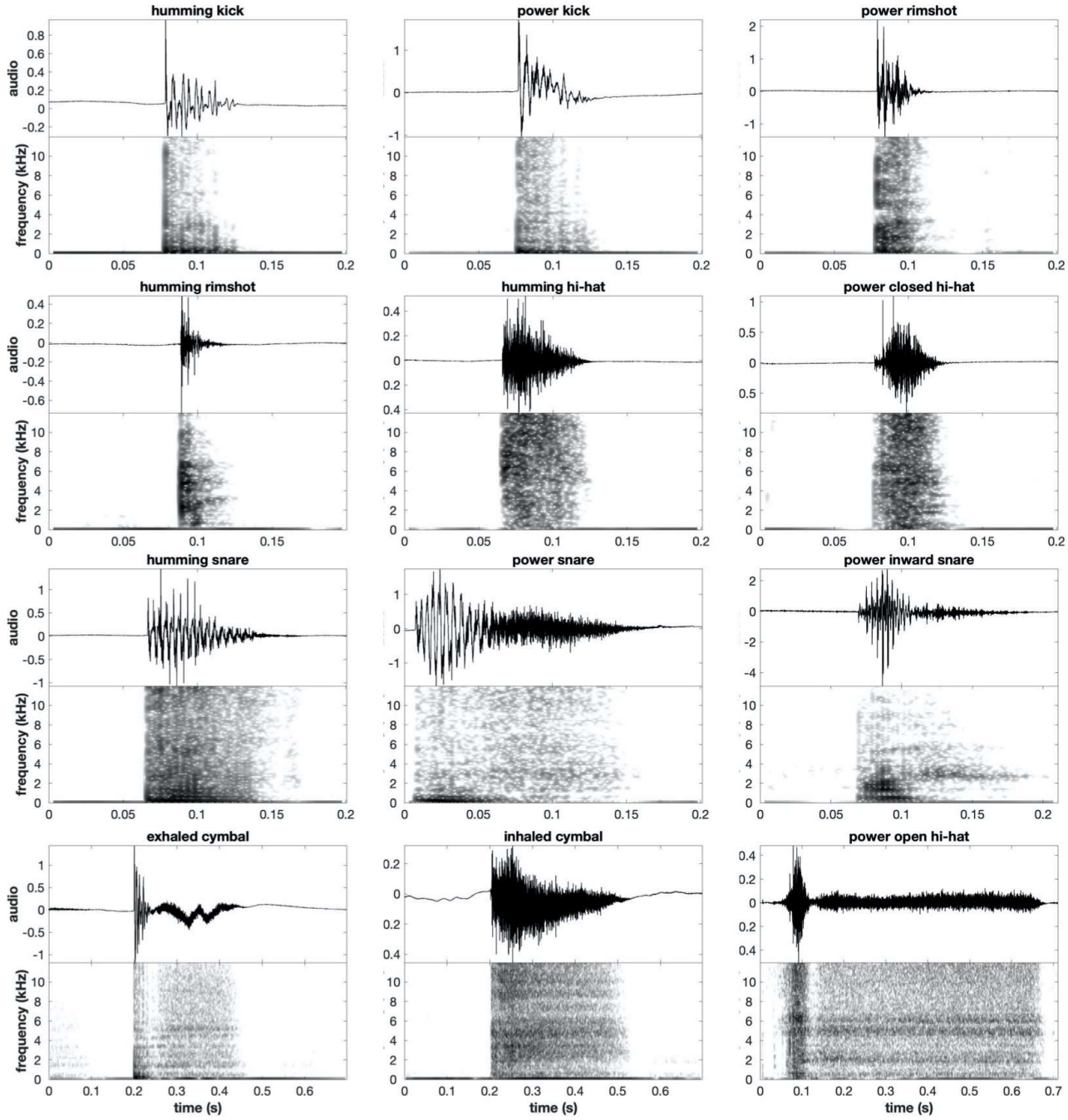


Figure 4.2: Audio waveforms and spectrograms of a representative token for each of the twelve HBB sounds. Spectrogram parameters: view range: 0-12 kHz; window length: 5 ms; dynamic range: 50 dB.

Most sounds have a duration shorter than 200 ms (Fig. 4.2 and Fig. 4.3). Only three sounds were associated with longer duration, ranging from 300 ms to 700 ms. Six sounds

are impulsive sounds, most often produced with a strong burst: humming and power kick, humming and power rimshot, humming and power closed hi-hat. The others are characterized by an impulse attack followed by a more or less protracted friction noise: power snare and inward snare, exhaled cymbal and inhaled cymbal, power open hi-hat. Some sounds show a vibration component, either for the whole sound (humming kick and snare) or for the attack (power snare, inward snare, exhaled cymbal). The EGG signal does not show any signs of vocal-fold vibration (Fig. 4.6, 4.8 and 4.10), hence indicating that the vibratory source is located elsewhere than the glottis. The vibratory-source nature will be discussed in Section 4.4.6.

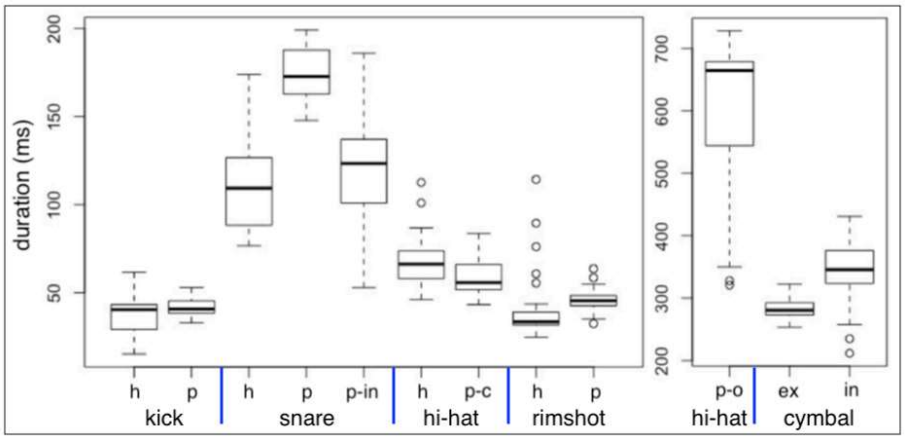


Figure 4.3: Distribution of duration for the twelve HBB sounds. Legend: h = humming; p = power; c = closed; o = open; in = inward/inhaled; ex = exhaled.

As shown in Fig. 4.3 and Table 4.2, HBB sound duration ranges from 37±11 ms for humming kick to 595±139 ms for power open hi-hat. Sound intensity ranges from 41±1 dB for the softest (power open hi-hat) to 60±1 dB for the loudest one (power snare), as shown in Fig. 4.4 and Table 4.2. Large variability among the sound realizations is clearly visible, especially for the power inward snare. The power version of all the effects is always produced at a higher intensity than the humming ones. The difference in intensity between power and humming variants of the same sound category is significant for the kick (0.1973±0.0110, $p<0.001$), snare (0.1412±0.0144, $p<0.001$), hi-hat (0.1179±0.0145, $p<0.001$) and rimshot (0.2034±0.0133, $p<0.001$) effects.

The ANCOVA analysis shows that sound duration and vocal intensity do not correlate with each other in most cases, except for three sounds (humming rimshot, power inward snare, and power closed hi–hat). Sound duration negatively correlates with intensity for humming rimshot (-0.0845±0.01467, $p<0.001$) and power closed hi–hat (-0.0559 ±0.0114, $p<0.001$), whereas a positive yet weaker correlation is found for power inward snare (0.0151±0.0.0049, $p<0.05$).

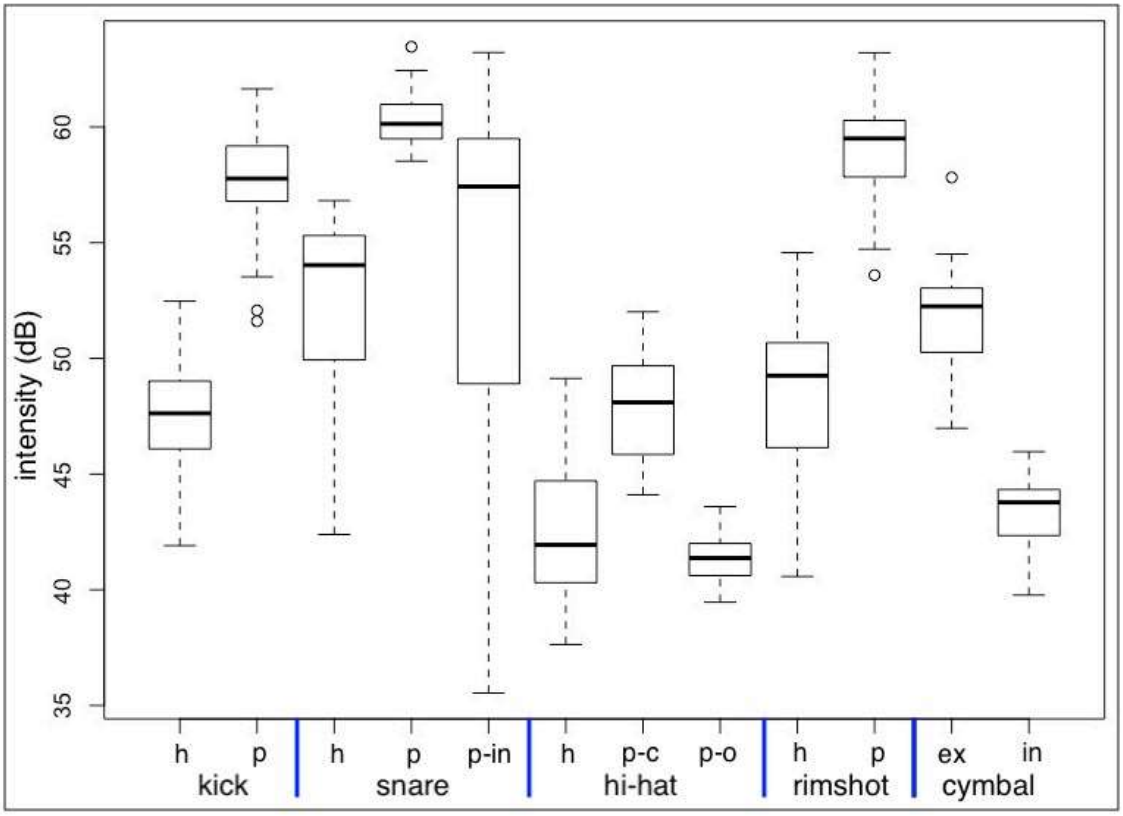


Figure 4.4: Distribution of vocal intensity for the twelve HBB sounds. Legend: h = humming; p = power; c = closed; o = open; in = inward/inhaled; ex = exhaled.

4.3.2 Articulatory characterization

Based on acoustic, EGG, video, respiratory and EMA data (see also multimedia material available online), the HBB drum sounds could be qualitatively interpreted as corresponding to a variety of articulatory and phonatory gestures, ranging from bilabial ejectives to lateral clicks, in addition to more common ones for French such as oral occlusives and fricatives. Some non-linguistic mechanisms were also observed. Fig. 4.5, 4.7 and 4.9 show the displacements of the lips and tongue sensors during 5 repetitions of the same sound. The trajectory of a representative gesture is highlighted in black. In general, coil trajectories are rather consistent over the repetitions, meaning that the articulatory pattern of each sound is stable.

Table 4.2: Mean and standard deviation (in brackets) of the sound duration and vocal intensity.

Sound	Duration (ms)	Intensity (dB)
humming kick	37 (11)	47 (3)
power kick	42 (5)	58 (2)
humming snare	112 (28)	52 (4)
power snare	174 (15)	60 (1)
power inward snare	120 (28)	53 (9)
humming hi-hat	68 (14)	43 (3)
power closed hi-hat	59 (11)	48 (2)
power open hi-hat	595 (139)	41 (1)
humming rimshot	42 (21)	49 (3)
power rimshot	46 (6)	59 (2)
exhaled cymbal	283 (17)	52 (2)
inhaled cymbal	339 (52)	43 (2)

4.3.2.1 Lip articulations

Five HBB sounds were produced with complete lip occlusion: humming and power kick, humming and power snare, and exhaled cymbal. The release is lateralized to the left portion of the lips, as evidenced by EMA and video data.

The lips undergo relatively large and fast protrusion displacements during the realization of the humming and power kicks, whereas their movements are smaller for the humming and power snares (Fig. 4.5 and 4.6) and the exhaled cymbal (Fig. 4.9). The tongue is very active in the articulation of both humming and power kicks and snares (Fig. 4.5): the tongue sensors display considerable movements along regular trajectories that are similar for humming kick and humming snare and for power kick and power snare, but differ between humming and power. For the humming sounds, the tongue is raised in the dorsal region against the palate, suggesting a back closure isolating the oral cavity from the rest of the vocal tract. The coil trajectories suggest a pushing action of the tongue from back to front and from right to left toward the point at the lips where the occlusion is released.

RIP data (Fig. 4.6) show that humming sound production takes place during both inhalation (increasing VR values) and exhalation (decreasing VR values), suggesting that sound production and breathing are dissociated. This supports the hypothesis that the airflow used in producing the sound is non-pulmonic, originating in the oral cavity. The articulatory pattern of the tongue suggests that it is lingual egressive. The realization of

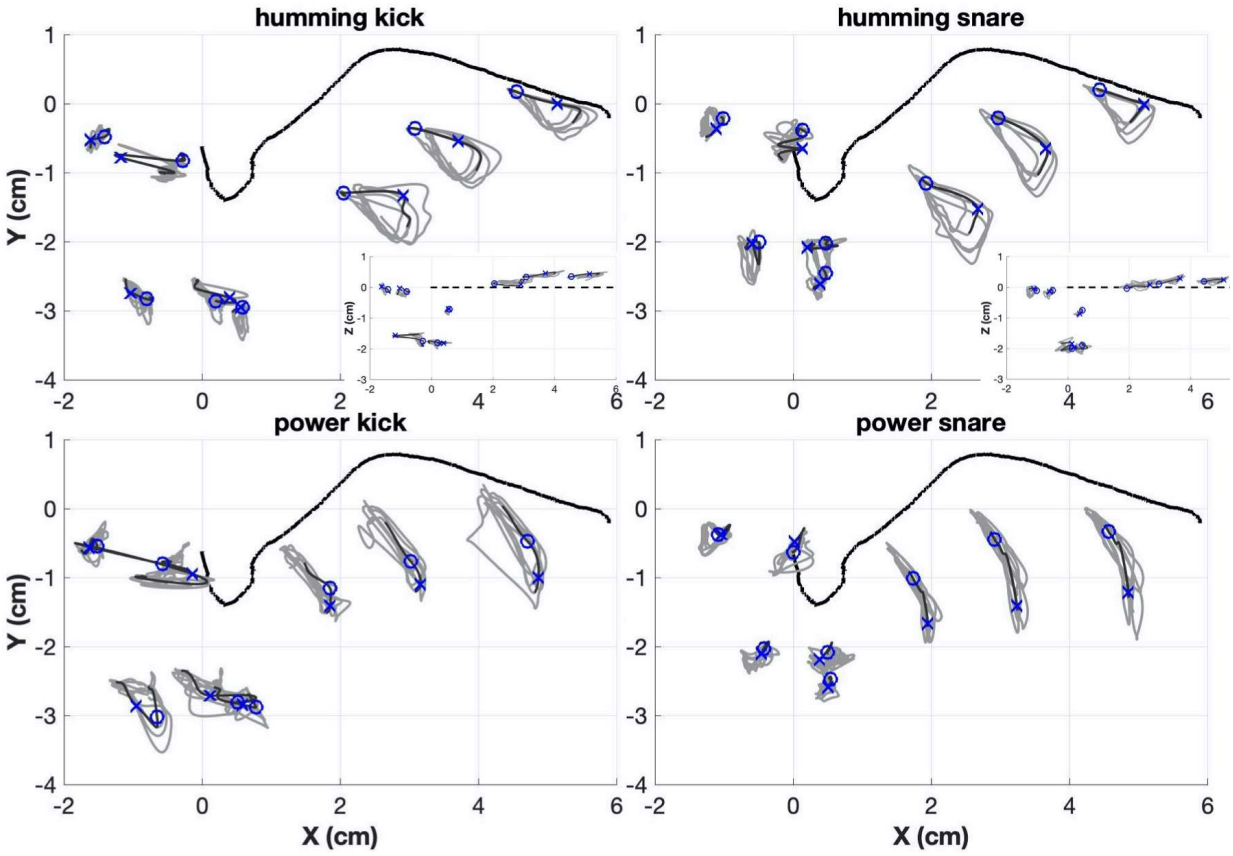


Figure 4.5: Sagittal (XY) and transversal (XZ) views of trajectories for 5 repetitions of kick sounds (humming/power) and snare ones (humming/power). Displayed coils: four lip coils, three tongue coils, jaw coil (see Fig. 3.2). Solid and dotted black lines: trace of the palate on the mid-sagittal plane. Black segment: trajectory of a representative token (same as Fig. 4.2). Grey lines: trajectories of the 2 tokens preceding and the 2 tokens following the representative token. Cross: start of sound. Circle: end of sound. Animation is available online as supplementary material.

the power sounds is achieved with a flatter tongue that moves from an overall lower to a higher (almost by 2 cm) position in the oral cavity. A laryngeal elevation is evidenced on the video. This movement is probably due to the use of an ejective mechanism. The shorter duration of the power kick sound with respect to the power snare one does not seem to reduce the overall tongue vertical displacement by much. Decreasing ventilatory volume (VR) values during sound production indicate that the airstream mechanism is egressive for both sounds (Fig. 4.6). For the power snare, the fricative portion of the sound (Fig. 4.2 and 4.6) is likely produced with a pulmonic egressive airstream. Video and acoustic data show that the stricture of close approximation related to this friction is created between the left portion of the lower lip and the upper teeth. In the exhaled cymbal (Fig. 4.9), the tongue, although moving slightly from a lower to a higher position during

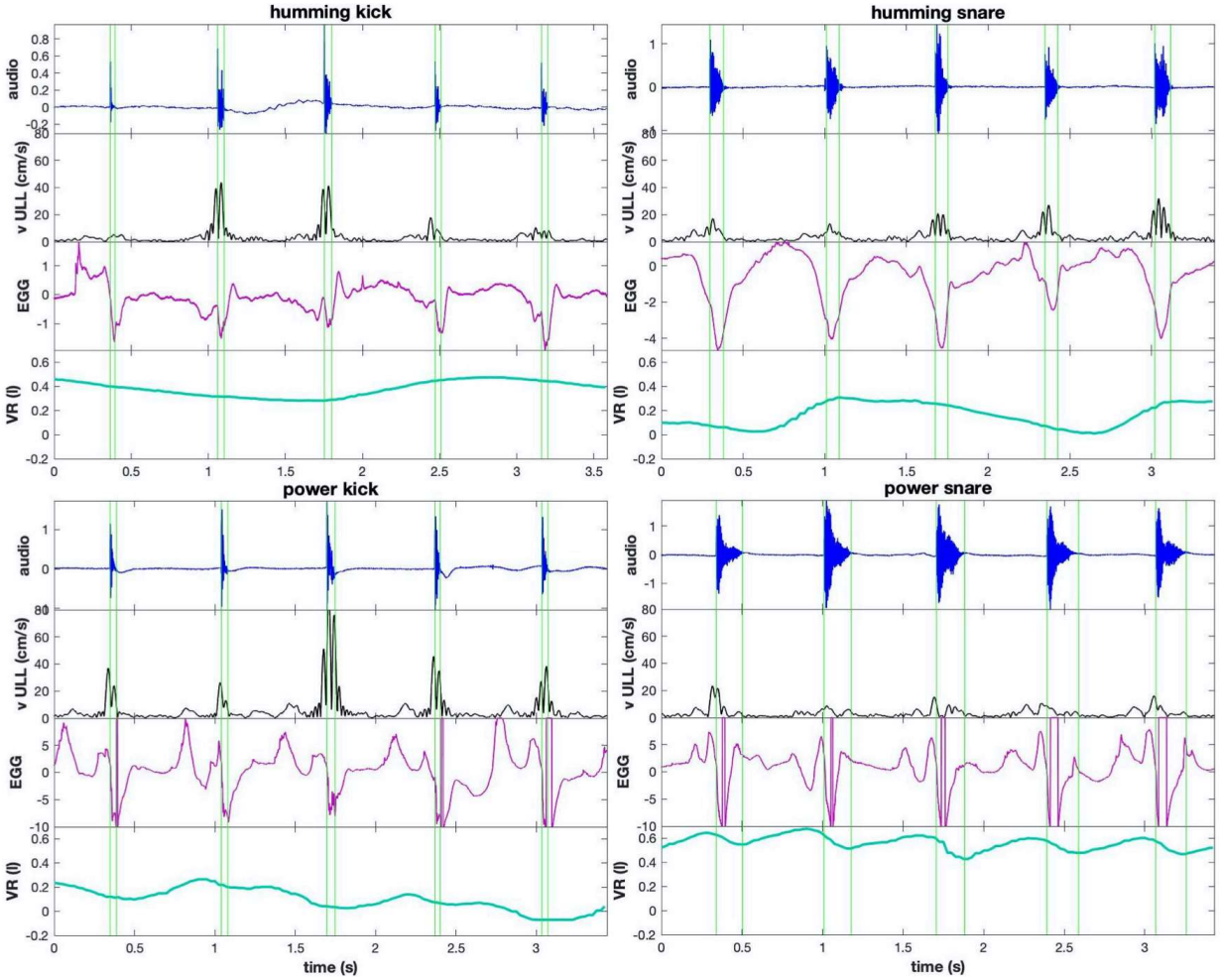


Figure 4.6: Synchronized audio, lip-coil speed (vULL), EGG, and RIP data (ventilatory volume VR) of five repetitions of kicks (humming/power) and snares (humming/power) (same as Fig. 4.2 and 4.5).

sound production, especially its posterior portion, assumes an almost horizontal position, revealing a laminar articulation of the fricative portion of the sound (Fig. 4.2). As for the power snare, the airstream is egressive (decreasing VR values) (Fig. 4.10). Video data show slight larynx elevation, suggesting the use of an ejective articulation for the bilabial.

4.3.2.2 Anterior tongue articulations

Four sounds were produced with complete occlusion of the vocal tract in the alveolar or post-alveolar region: humming hi-hat, power closed and open hi-hat, inhaled cymbal. Different tongue positions and the use of different airstream mechanisms differentiate the

realization of these sounds.

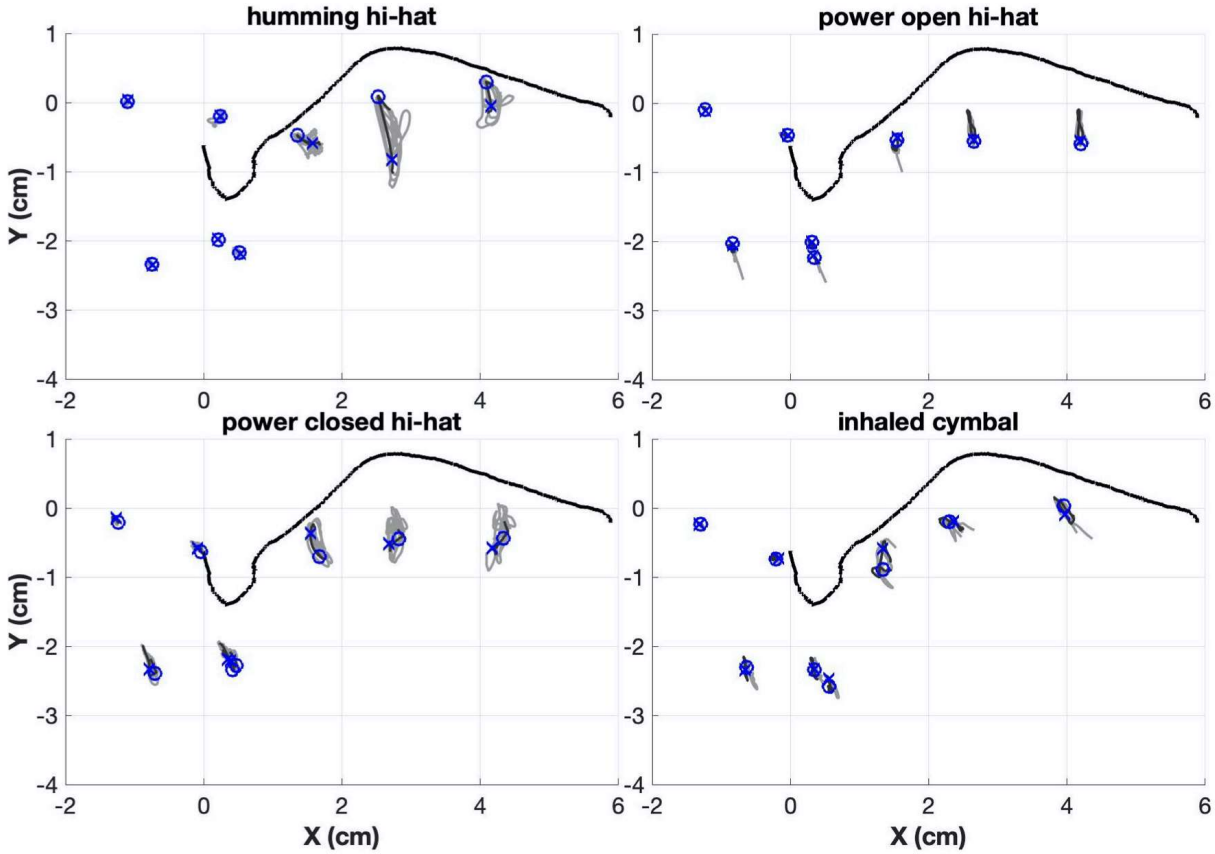


Figure 4.7: Sagittal (XY) views of trajectories of 5 repetitions of power closed hi-hat, power open hi-hat, humming hi-hat, inhaled cymbal. Legend: see Fig. 4.5.

The articulatory data for humming hi-hat (Fig. 4.7) show that the tongue forms a cavity in the mid-region, suggesting that a pocket of air is trapped between the alveolar/post-alveolar and dorsal regions. The mid-region of the tongue is then rapidly pushed upward (Fig. 4.8) during sound production, suggesting that the oral airflow is indeed generated by a pushing action of the tongue. RIP data (Fig. 4.8) show that sound production takes place during both exhalation (decreasing VR values) and inhalation (increasing VR values). This is evidence for the use of a non-pulmonic airflow that allows some dissociation between sound production and ventilation. The combination of the articulatory pattern and the breathing behavior suggests that this gesture is produced via a lingual egressive airstream mechanism. The lip coils hardly move, meaning that the lips are not active in the articulation of this sound. The posterior seal may take place in the velar region, further back than the DORS coil. The anterior seal may take place in the alveolar or post-alveolar region and may be apical rather than laminal. This would explain the almost horizontal trajectory of TIP coil during sound production.

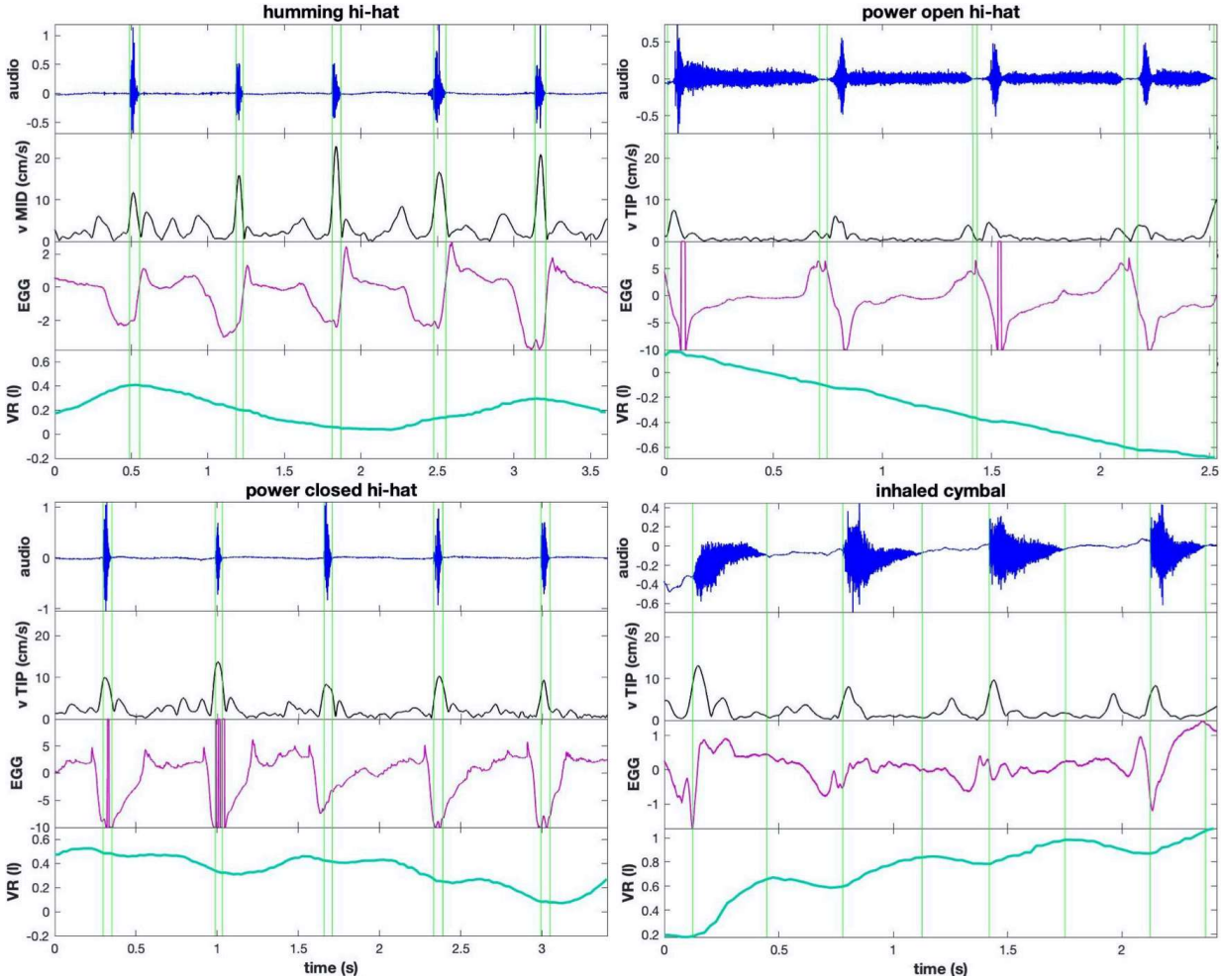


Figure 4.8: Synchronized audio, speed, EGG, and RIP data of five repetitions of power closed hi-hat, power open hi-hat, humming hi-hat, inhaled cymbal (same as Fig. 4.2 and 4.7).

The articulatory movements of the power closed hi-hat are quite subtle and mainly restrained to the tip region (Fig. 4.7), especially during sound production, while the tongue assumes a generally flat position in the middle of the oral cavity. The vertical movements of the tongue, especially in its mid and dorsal regions, may be due to an upward movement of the larynx evidenced on the video and likely related to an ejective mechanism.

The power open hi-hat (Fig. 4.7) is produced similarly to the closed version, but the alveolar occlusive is followed by a laminal constriction. Again, the vertical displacement of the tongue during the first part of the sound production may be related to the upward movements of the larynx (Fig. 4.7 and supplementary material). The airstream is clearly egressive (decreasing VR values), likely glottal at first, then pulmonic.

The inhaled cymbal is realized with the tongue in an arched and higher position than the other sounds (Fig. 4.7). The airstream used is pulmonic ingressive (increasing VR values during sound production) (Fig. 4.8).

4.3.2.3 Posterior tongue articulations

Three sounds were articulated with complete occlusion of the vocal tract in the posterior region of the oral cavity: power inward snare, humming rimshot, and power rimshot.

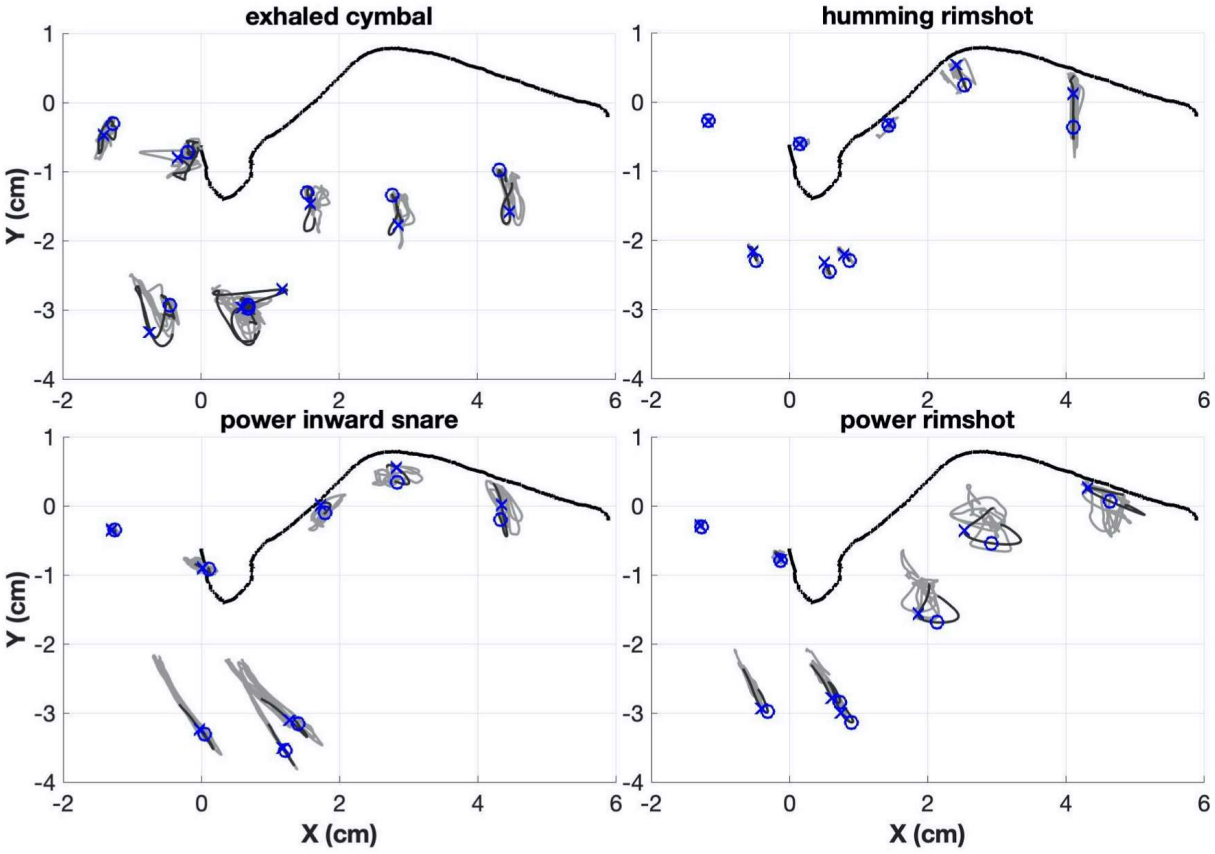


Figure 4.9: Sagittal (XY) views of trajectories of 5 repetitions of humming rimshot, power rimshot, exhaled cymbal, power inward snare. Legend: see Fig. 4.5.

The front portion of the tongue is held against the hard palate in the production of the power inward snare and humming rimshot while the occlusion is released in the dorsal region (Fig. 4.9).

Sound production during both exhalation (decreasing VR values) and inhalation (increasing VR values) (Fig. 4.10) indicates that the airstream of the humming rimshot is non-pulmonic. The aggregation of articulatory (Fig. 4.9), ventilatory (Fig. 4.10), and acoustic

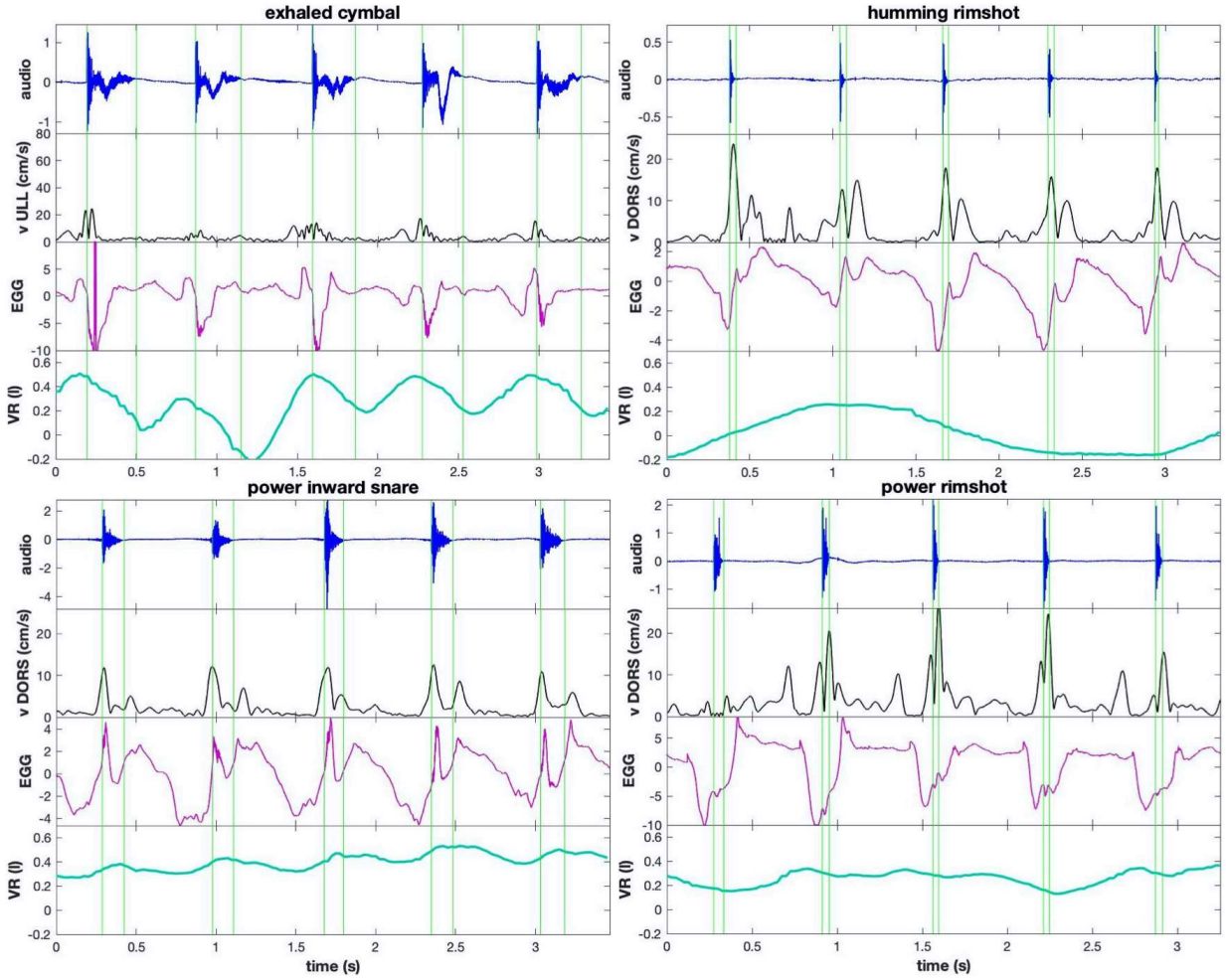


Figure 4.10: Synchronized audio, speed, EGG, and RIP data of five repetitions of humming rimshot, power rimshot, exhaled cymbal, power inward snare (same as Fig. 4.2 and 4.9).

(Fig. 4.2 and supplementary material available online) data implies that the airstream is lingual ingressive (or velaric).

The power inward snare shows a downward motion of the jaw and lower lip. Increasing VR values during sound production (Fig.4.10) suggest that the airstream is pulmonic ingressive. In the power rimshot, only the posterior part of the tongue is in contact with the palate. Before the burst, the motion of the sensors suggests that the tongue is pushed upward and forward, while the occlusion is being held. When the burst occurs, the tongue dorsal region (DORS coil) reaches its highest and most advanced position while the jaw is lowered together with the lower lip (JAW, LLL and LLM coils). Systematic decrease of respiratory volume during sound production (Fig. 4.10) indicates that the airstream of the power rimshot is egressive. Upward movements of the larynx suggest the use of an ejective mechanism.

4.3.3 Articulatory dynamics

The analysis of maximum speed distribution is presented in Fig. 4.11. Lips are the articulators that reach the highest values of speed, especially on their left side (ULL and LLL coils). Power variants show faster moves than humming ones. In the power kick, the left upper lip has an average maximum speed of 45 cm/s, but it can reach maximum velocities as high as 90 cm/s. Humming and power snares both involve a bilabial occlusive, however, the order of magnitude of lip speed is smaller (15-17 cm/s for the upper left lip for both variants) than the kicks, possibly because the lips are still engaged in a stricture of close approximation after the release of the occlusion.

The data show that the tongue is almost always involved in the articulation of the explored HBB sounds, either as the main articulator or accompanying the lip dynamics. However, it never reaches the highest speed values of the lips. Our analyses point out that the tongue, either as a whole or in part, is the main articulator for the production of both the humming and the power variant of the hi-hat and rimshot effects, the power inward snare as well as the inhaled cymbal. The regions of the tongue that reach the highest velocities typically match the main place of articulation, *i.e.* where the occlusion is released. However, in the humming hi-hat, the mid-portion of the tongue appears to be the fastest moving articulator, moving at an average maximum speed of about 20 cm/s. As discussed in the previous section, this is likely the place where the airflow is generated and not the place where the anterior occlusion is released.

The sounds for which the tongue is not the main articulator are both the humming and the power variants of kick and snare, as well as the exhaled cymbal. In these cases, a general tendency seems to emerge that the tongue moves as a whole, with all three regions showing comparable average maximum velocities.

The analysis of speed distribution demonstrates limited dynamics for the jaw. This articulator almost never reaches high speeds, moving at an average maximum speed of approximately 5 cm/s across all the examined sounds. The jaw dynamics seem to be quite independent of the dynamics of the left lower lip (LLL coil) in all bilabial effects. The statistical analysis shows that the JAW coil reaches significantly lower maximum speed values than the LLL coil in all these sounds (humming kick: -1.5258±0.0672, $p<0.001$; power kick: -1.4389±0.0513, $p<0.001$; humming snare: -1.1822±0.0601, $p<0.001$; power snare:-0.9178±0.0519, $p<0.001$; exhaled cymbal: -1.4924±0.1075, $p<0.001$).

Only in the articulation of two HBB sounds, *i.e.* power rimshot and power inward snare, the jaw moves more quickly, reaching approximately 10 cm/s for the former and slightly less than 15 cm/s for the latter. In both cases, the jaw dynamics possibly accompanies the lower lip dynamics, as the two articulators (JAW, LML and LLL coils) on average show the same maximum velocities.

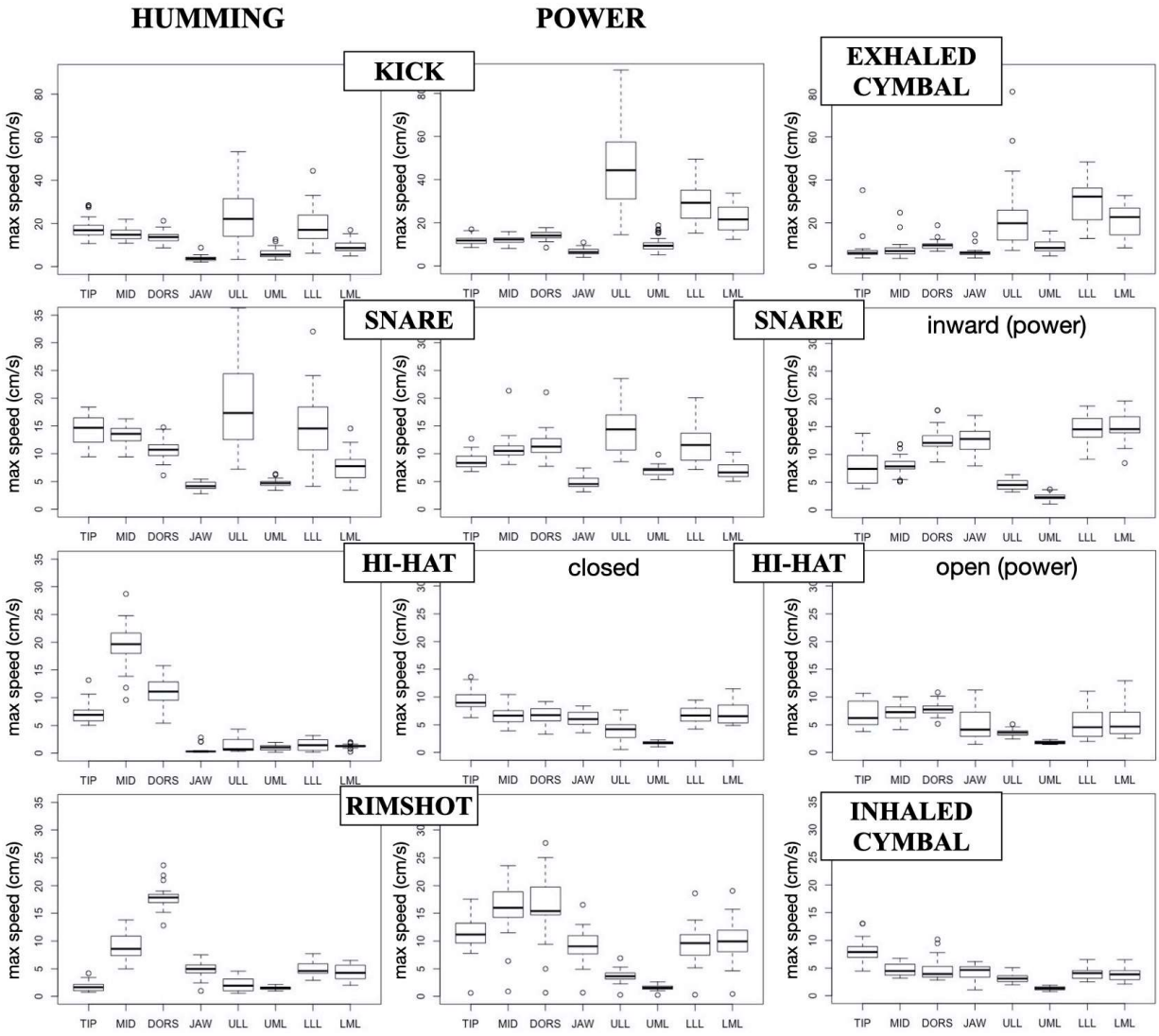


Figure 4.11: Maximum speed distribution (in cm/s) of the coils for the twelve HBB sounds. Left column: humming variants; center and right column: power variants and cymbals. Note that the first row of panels has a wider y-axis scale, because of faster lip movements for kick and exhaled cymbal sounds.

4.3.4 Phonetic description

A phonetic annotation was performed using the IPA alphabet. The results are presented in Table 4.3.

In general, either the symbols utilized do not belong to French or, if they are present in French, some diacritics were needed, because of the occurrence of perceptible phonetic effects or modifications. A symbol was assigned to each sound. Especially the non-speechlike

Table 4.3: Phonetic characterization and brief articulatory description of the HBB sounds.

Sound	IPA	Description				Articulation
		voicing	airstream	place	manner	
humming kick	[ʘ͡β̥ˡ↑]	voiceless	lingual egressive	lateral bilabial	stop	double
		voiceless	lingual egressive	lateral bilabial	trill	
humming snare	[ʘ͡β̥ˡːˑ↑]	voiceless	lingual egressive	lateral bilabial	stop	double
		voiceless	lingual egressive	lateral bilabial	trill	
power kick	[pʼˡ]	voiceless	glottalic egressive	lateral bilabial	stop	simple
power snare	[p͡fˡːˑ]	voiceless	glottalic egressive	lateral bilabial	stop	double
		voiceless	pulmonic egressive	lateral labio–dental	fricative	
exhaled cymbal	[β̥ʼˡs̻ːˑ]	voiceless	glottalic egressive	lateral bilabial	trill	double
		voiceless	pulmonic egressive	laminal	fricative	
humming hi–hat	[|↑]	voiceless	lingual egressive	alveolar	stop	simple
power closed hi–hat	[t̻ʼ]	voiceless	glottalic egressive	alveolar	stop	simple
power open hi–hat	[t̻ʼs̻ːˑ]	voiceless	glottalic egressive	alveolar	stop	double
		voiceless	pulmonic egressive	alveolar	fricative	
inhaled cymbal	[t̻s̻ːˑ↓]	voiceless	pulmonic ingressive	alveolar	stop	double
		voiceless	pulmonic ingressive	alveolar	fricative	
humming rimshot	[‖]	voiceless	lingual ingressive	lateral	stop	simple
power rimshot	[kʼ]	voiceless	glottalic egressive	velar	stop	simple
power inward snare	[k͡ɬ↓]	voiceless	pulmonic ingressive	velar	stop	double
		voiceless	pulmonic ingressive	lateral	fricative	

articulatory and airstream mechanisms required the use of diacritics. Neither the IPA nor the extIPA (Ball et al., 2018) provide a notation for lingual egressive articulations. Hence, the symbol for the corresponding click (always ingressive in speech) was used in combination with the symbol for an egressive airflow. The vibratory aspects revealed by the acoustic investigation (Section 4.3.1) are annotated as a voiceless bilabial trill (β̥). Due to the lack of a symbol for a lingual egressive mechanism, the difference in airstream mechanism presented in Section 4.3.2 cannot be reported in this table. Further, the frequency of lip vibration of these sounds seems higher than that of speech bilabial trills.

Some sounds presented similarities with French phonemes: bilabial, alveolar and velar stops, labiodental and alveolar fricatives. However substantial features differentiate the HBB sounds from the French phonemes. The power kick is similar to the French [p] in that it is a bilabial stop. It is however ejective and lateralized. Similarly, the power closed hi–hat and the power rimshot are similar to the French [t] and [k], except for the airstream mechanism.

4.4 Discussion

4.4.1 Feasibility and suitability of multimodal synchronized physiological measurements in HBB

Although limited to one subject (as is often the case for HBB, *e.g.* Blaylock et al., 2017; Proctor et al., 2013), the present study suggests that the recording of multimodal (EMA, EGG, RIP, audio, video) and synchronized data is compatible with HBB production and paramount in the exploration and understanding of the production mechanisms of this peculiar vocal art.

The beatboxer was able to produce more than one hour of sounds with the coils firmly attached to the lips and tongue. The coil wires were uncomfortable for him at first, but he got used to them and managed to produce all the HBB sounds in the protocol. The measurements consisted of three-dimensional articulatory movements. Being able to compute tangential speed with all three spatial components (x, y, z) was particularly relevant for lip dynamics in HBB, which presented several lateral articulations such as lateralization of occlusion release.

4.4.2 Boxemes, distinct sound units

The acoustic data outlined different spectral signatures for every sound. These differences were such that an unsupervised classifier was able to automatically detect each sound and correctly assign it to a category in agreement with those provided by the beatboxer. The articulatory and ventilatory data also showed different behavior that distinguishes each sound from the others, in terms of place and/or manner of articulation, and airstream mechanism. Our results indicate that each one of the twelve HBB drum sounds investigated in this study was substantially different from the others, supporting the idea that they make sense as distinct sound units. We propose that these sound units be called *boxemes*, by analogy with speech phonemes. They constitute the building blocks of a HBB musical phrase. Considering HBB as a musical language structured similarly to human speech calls for future research that goes far beyond the present study.

The few studies that have proposed an IPA transcription of speech-like HBB sounds show some degree of agreement with the transcription proposed in the present investigation. Kicks that correspond to power kick in this study often involve bilabial ejectives [p'] (Blaylock et al., 2017; Patil et al., 2017; Proctor et al., 2013), snares corresponding to power snare are a double articulation of a bilabial ejective and labiodental fricative [p'f] or [pf'] (Blaylock et al., 2017; Patil et al., 2017; Proctor et al., 2013), hi-hats corresponding to

power closed and open hi-hat often involve an alveolar stop [t] or [ts] (Blaylock et al., 2017; Patil et al., 2017; Proctor et al., 2013), rimshots corresponding to power rimshot are often velar stops [k] or [k'] (Proctor et al., 2013). Even though Blaylock et al. (2017) do not provide an IPA transcription, similarities in non-linguistic articulations can be found between power inward snare and inward K, humming kick and lip pop, humming hi-hat and forced hi-hat. Such an agreement suggests that similar acoustic and/or articulatory strategies are used for the same sound among different beatboxers, regardless of the beatboxer's native language.

4.4.3 Complex articulatory behaviors

Our data demonstrated a variety of articulatory gestures, many of which elicited labial dynamics. Lingual dynamics was also much elicited, both when the tongue was the main articulator and when accompanying lip dynamics. This suggests complex tongue-lip synergies. On the contrary, jaw dynamics was often limited in our corpus of HBB drum sounds, possibly due to the absence of vocalic sounds.

The investigated sounds seem to be produced on two different time scales: 9 sounds were short, generally not exceeding 200 ms, 3 were longer, up to 750 ms. However, even the shorter sounds could be produced as a double articulation of a plosive attack, generally due to the release of a complete occlusion, followed by a friction noise.

Our data showed the use of quite a wide variety of manners of articulation. Despite the small number of HBB drum sounds explored, ejectives, clicks, stops and fricatives were observed. Most of the produced sounds did not belong to the phonology of French, the language spoken by our subject. Some are found in other world's languages (Ladefoged & Maddieson, 1996), others have never been attested in any language.

4.4.4 Mastering pulmonic and non-pulmonic airstreams

In agreement with the existing literature (Blaylock et al., 2017; Proctor et al., 2013), both egressive and ingressive airstreams were observed. In some cases, the opposite airstream was used as compared to what is generally observed for the speech counterparts of the same sounds. This occurred mainly in the articulation of stop and fricative sounds, where a pulmonic ingressive airstream could be used. Stowell and Plumbley (2010) also described the use in HBB of a given sound produced with both pulmonic ingressive and egressive airstream in the case of oral stops.

All the humming sounds were produced via a lingual ingressive (velaric) or egressive

airstream. The latter has already been described by Blaylock et al. (2017), but it has never been observed in speech so far. A lingual airflow initiation grants some dissociation between sound production and articulation. This allows the beatboxer to perform multiple actions at the same time, such as breathing or producing a melodic line through the nose without being silent.

However, this has a cost in terms of intensity, as humming variants were always significantly quieter than their power counterparts.

4.4.5 Evidence for ejective productions

Our data argue in favor of an ejective production of all the non-humming sounds that imply a bilabial occlusion (e.g. kick and snare effects), or alveolar occlusion with egressive airstream (e.g. power closed hi-hat). An ejective production of the power rimshot could not be precluded as a possibility. As mentioned in Section 2.2, the use of ejectives in HBB was already attested (even though not systematically employed by all beatboxers, Patil et al., 2017) by a few articulatory studies that exploited different imaging techniques (video-fiberscopy De Torcy et al., 2014; Dehais Underdown et al., 2019; Sapthavee et al., 2014, MRI Blaylock et al., 2017; Patil et al., 2017; Proctor et al., 2013). In particular these studies characterize several kick and snare sounds as ejectives, when produced as bilabial occlusives, closed hi–hats as alveolar ejectives as well as a rimshot sound as a velar ejective. Proctor et al. (2013) characterized three kick sounds as stiff ejectives, with different amounts of lingual retraction during laryngeal lowering and a different final lingual posture. They also suggested that tongue and larynx may be used in concert to produce a more effective pushing action. Our data support this hypothesis of a lingual action in the articulation of ejective sounds.

4.4.6 Vibration and lateralization

In some cases, acoustic data revealed a clear vibratory pattern that did not originate from glottal vibration as attested by the EGG signal. The combination of acoustic, articulatory and video data suggests that the vibration was produced at the place of occlusion, namely the lip area for the humming kick and snare, power snare and exhaled cymbal, and possibly the lateral rim of the tongue for the power inward snare.

All the bilabial sounds were laterally released on the left side of the lips. This consistent lateralization may be explained by the fact that beatboxers need to control lip tension in order to produce self-oscillation at adequate vibratory frequency (Stowell & Plumbley, 2010). Shortening the lip portion that can vibrate may provide a better control and the

possibility to produce vibrations at higher frequencies. The resulting effect is reminiscent of the way vocal folds are controlled to modify f0.

Furthermore, the lateralization of the bilabial occlusion release seemed to affect the articulation of the following fricative at least in the case of the power snare. Here, the labiodental fricative was also articulated on the left.

4.4.7 HBB sound annotation

Using IPA to annotate HBB sounds was not straightforward, in agreement with Blaylock et al. (2017). Some basic HBB sounds may stem from speech sounds (e.g. classic kick, or power kick according to our beatboxer's terminology). They may share the same mechanisms, as suggested by Proctor et al. (2013), but they are substantially modified to induce a non-linguistic or para-linguistic connotation. Further, our data displayed the use of sources of vibration other than the glottal one, suggesting that the simple distinction between voiced and voiceless sounds is not sufficient in HBB to fully characterize the acoustic production. Moreover, even if the international HBB community shares a considerable amount of coded sounds, a prerogative of each beatboxer is to experiment with their own vocal instrument to create new sounds, never produced before and more and more difficult to articulate. As a consequence, a much more subtle and adapted notation system is needed in order to capture the acoustic and articulatory richness of HBB production. An articulatory-based pictographic writing system (Vocal Grammatics) seems promising and has recently been used for beatbox–sound automatic recognition purpose (Evain et al., 2020).

4.5 Conclusion and perspectives

Acoustic, articulatory and ventilatory properties of twelve different HBB drum sounds were investigated on a French beatboxer. Electromagnetic articulography, an experimental technique widely used in speech research, was successfully used to capture the articulatory behavior. It was combined with acoustic measurements, electroglottography and respiratory inductance plethysmography to get a deeper understanding of articulatory and airstream mechanisms underlying these complex vocal sound productions.

In agreement with the existing literature, a wide variety of articulatory gestures were observed, most of which do not belong to the phonology of the beatboxer's language, nor to any known phonology. Our data revealed the use of multiple airstream mechanisms, the possibility of dissociating breathing and sound production, a pronounced labial dynamics,

or a lingual dynamics that accompanies the labial dynamics when the principal articulator is not the tongue.

The notion of *boxeme* has been suggested, as building blocks of human beatboxing considered as a musical language. This, however, calls for further research.

This investigation was conducted on a single beatboxer. Analysis of HBB articulatory behavior from multiple beatboxers with several training levels is needed, in order to generalize these findings and relate them to the HBB level of practice.

Chapter 5

Beatboxing, is it speaking?

Contents

This chapter characterizes the similarities and differences in articulatory strategy and breathing behavior of three boxemes (kick, hi-hat, rimshot) and the three French consonants **p**, **t**, **k**[1].

[1]This chapter is based on work presented at three conferences (Journées d'Etudes sur la Parole 2020, 2022, International Seminar on Speech Production 2020).
Sources: Paroni, A., Henrich Bernardoni, N., Savariaux, C., Baraduc, P., and Lœvenbruck, H. (2020). Beatboxer, est-ce parler ? Ce que nous en dit l'étude de la dynamique articulatoire d'un beatboxeur. Available at: https://hal.archives-ouvertes.fr/hal-02798574v1
Paroni, A., Henrich Bernardoni, N., Savariaux, C., Baraduc, P., Calabrese, P., and Lœvenbruck, H. (2020). Beatboxing, is it talking? *Proc. 12th International Seminar on Speech Production*, pp. 238–241. Available at: https://issp2020.yale.edu/ProcISSP2020.pdf
Paroni, A., Henrich Bernardoni, N., Lœvenbruck, H., Gerber, S., Baraduc, P., and Savariaux, C. (2022). Etude comparative acoustique et articulatoire de la plosion entre parole et beatbox. *Accepted.*

5.1 From "boots and cats" to P ts K ts

Stop or plosive consonants /p, t, k/ are among the most commonly-found phonemes in the phonological inventory of the world's languages (Maddieson & Disner, 1984). These sounds are produced via complete occlusion at different places along the VT (Sec. 1.2.1). In addition to their linguistic role, they also fulfil a non-linguistic role. Vocal percussion is practiced in many vocal cultures around the world (Patel & Iversen, 2003). Syllables with plosive consonants are used in instrumental practice, for playing wind instruments and percussion. They support vocal practice as well, as is the case with skat and konnakol (Arleo, 1999; Atherton, 2007). Some studies have highlighted articulatory differences in the stops used in instrumental practice depending on the player's first language (Heyne & Derrick, 2014, 2015; Heyne et al., 2019; Lamkin, 2005). Plosive sounds are also commonly used in HBB. HBB learning frequently begins with training on speech plosives, syllables or sentences. For instance, kicks, the imitations of the kick drum, are commonly learned from a [p] or [b] consonant, the imitations of the hi-hats can stem from a [t] or [ts], and the imitations of the rimshot technique performed on the snare drum are often based on a [k]. Simple *beats* are often learned based on spoken sentences. Perhaps the most famous sentence used worldwide is "boots and cats". It provides the basis to learn the most simple beat in four four rhythm[2]: kick, hi-hat, rimshot, hi-hat, or in SBN **P t K t**. Usually, indications are given to emphasize the consonants and firstly devoice, then skip the vowels completely. However, the exact adaptations that occur are unexplored from a scientific standpoint. Multiple studies (Blaylock et al., 2017; De Torcy et al., 2014; Dehais Underdown et al., 2019; Patil et al., 2017; Proctor et al., 2013; Sapthavee et al., 2014) have shown that often these sounds are glottalic egressive. i.e., produced via a non-pulmonic airstream. A study (Dehais Underdown et al., 2019) provided details on the physiology of a beatboxer's kick, with acoustic and aerodynamic data. However, comparative studies between HBB and speech are rare in the literature. The research question tackling the similarities and differences in production mechanisms of comparable speech and HBB sounds remains very much open: how does "boots and cats" become **P t K t**? Further, in *beatrhyming* speech and boxemes are combined: speech sounds or even entire words of the lyrics are replaced with HBB sounds to give the impression that both the lyrics and a rhythmic line are produced at the same time. Segment replacements seem to be fulfilled by sounds with transient noise, e.g., kicks and snares, whereas sustained sounds are more likely to be chosen for word replacements, and affricates (that have both properties) may be used for both purposes (Fukuda et al., 2022). As the authors suggest, this indicate that beatboxers resort to some phonetic and phonological knowledge of speech when beatrhyming, but it is not yet understood to what extent.

[2]A four four beat is the simplest rhythm of modern music. Its simplest form is: 1 kick, 1 snare, in first and third position, and then hi-hats or cymbals in the less prominent positions.

Further, breath supply is necessary for phonation. Breathing constitutes the support to oral communication and human vocalisation in general. The speaker, and more so the singer, are constantly in quest for the control between their produced sounds and their breath support. In general, and particularly in French and English, the air used in speech is mostly that coming from the lungs. This pulmonic airstream regulates subglottal pressure, and therefore, breath control is an important aspect of voice production (Leanderson & Sundberg, 1988; Titze & Martin, 1998). In order to produce speech, breathing mechanisms and hence subglottal pressure must be finely controlled to meet the needs of human communicative tasks (MacLarnon & Hewitt, 1999), as well as the equilibrium between active and passive (recoil) forces in the breathing system (Huber & Stathopoulos, 2015). Because it vastly relies on subglottal pressure, linguistic speech production usually takes place during the exhalation phase. In order not to disrupt the linguistic communication, air intake must be relatively rapid and happen at linguistically relevant times (Huber & Stathopoulos, 2015; MacLarnon & Hewitt, 1999). This holds true also in the case of artistic voice production such as singing: due to rhythmical constraints, time of inspiration decreases with respect to quiet breathing (Salomoni et al., 2016) and sound production generally occurs during the exhalation phase, where subglottal pressure is finely controlled via expert movements of the abdomen (Thorpe et al., 2001; Traser et al., 2017) and respiration accessory muscles (Pettersen, 2005). Despite following a common pattern of inhalation followed by exhalation, the breathing pattern is specific to an individual and is consistent over time (Benchetrit et al., 1989) and across different conditions (Benchetrit, 2000). Research has shown that, while breath management changes based on the demands of sound production (i.e., during artistic voice production such as singing as compared to speech) and expertise (neophites as compared to professionals), a general pattern of rapid, silent inhalation followed by a longer exhalation providing subglottal pressure necessary for phonation has been depicted. To the best of our knowledge, no published study is available so far that investigates breathing behavior in the case of HBB vocal production. As previously mentioned, some studies have found that non pulmonic airstreams are often used even for basic vocal drum sounds such as kick, hi-hat, and rimshot (Blaylock et al., 2017; De Torcy et al., 2014; Patil et al., 2017; Proctor et al., 2013). Our research presented in the previous chapter points in this direction as well (Ch. 4). Recent data attested the use of all 6 possible airstream initiation mechanisms for the production of isolated HBB sounds (Dehais-Underdown et al., 2021). However, no data is available as to how they interact over a full musical phrase. Furthermore, experienced beatboxers can produce extremely long and technically demanding beats without audible pauses for air intake. This raises the question of what breathing strategies they use to achieve extremely long phonatory phases and how they compare to those used in speech and other singing styles.

This chapter investigates the production of three plosive consonants of speech [p, t, k] compared to three HBB consonantal sounds with similar places of articulation, kick (**P**), hi-hat (**t**), and rimshot (**K**), focusing on the acoustic characteristics of the sounds, the

articulatory and breathing behaviors.

5.2 Material and methods

The results presented in this chapter are drawn from C1.III and C2.I. Preliminary observations (Sec. 5.3) are based on C1.III and more evidence (Sec. 5.4) was gathered on data from C2.I.

Table 5.3 summarizes the participant, the items, and the techniques relative to the data used to conduct preliminary observations (Sec. 5.3). For more details, see section 3.2.1.

Table 5.1: Visual summary of participant, items, and techniques employed.

Participants	Items	rep nb HBB	rep nb speech	Techniques
PS	"Boots" HBB and speech	23	16	EMA, RIP,
	"Cookies" HBB and speech	8	8	EGG,
	"Pâtes" HBB and speech	17	8	audio, video

The audio data were manually segmented and annotated in Praat (Boersma, 2006). The phonetic annotations were used to identify the moments when the acoustic bursts of the plosive sounds occurred. These temporal markers were saved in a TextGrid file. The articulatory and breathing data were then analyzed in MATLAB ("MathWorks: Bioinformatics Toolbox: User's Guide (R2018b)," 2018).

In order to visualize the prototypical trajectory and the variability of the articulatory movements, the median trajectory and interquartile range of the coils of interest were computed taking into account all occurrences of the same sound for the same production mode (HBB or speech), choosing as temporal reference the moment of the burst detected on the audio data. The median tangential velocity (time derivative of the space trajectory) and interquartile range were also calculated. The space exploration of the coils was also investigated by means of smooth maps or heat maps that visualize the density of positioning in space, i.e. how often a given coil is present in a given spatial point. The whole trajectories from beginning to end of HBB or speech production of the three sentences were used to produce a map. As a second step of the analysis, the two maps were superimposed to better visualize the intersecting regions.

Breathing data were analyzed in terms of evolution of reconstituted volume (VR) over time as well as the thoracic and abdominal components. Due to the small number of

repetitions available, central tendencies were calculated as median and interquartile range over a time normalized sentence.

The results presented in section 5.4 are drawn from C2.I from all 5 participants. Table 5.2 summarizes the participants, the items, and the techniques relative to the data used. For more details, see sec. 3.2.2.2.

Table 5.2: Visual summary of participants, items, and techniques employed.

Participants	Items	Techniques
S01	/pu/ – kick (**P**) /ti/ – hi-hat (**t**) /ka/ – rimshot (**K**) /putikati/ – **PtKt**	RIP, EGG, audio, video
S02-S05	/pu/ – kick (**P**) /ti/ – hi-hat (**t**) /ka/ – rimshot (**K**) /putikati/ – **PtKt**	EMA, RIP, EGG, audio, video

The audio, and electroglottographic data were manually segmented and annotated in Praat (Boersma, 2006). In particular, each consonantal sound was annotated by placing the onset boundary at the instant of the burst detected on the acoustic signal and the end boundary either (i) at the voicing onset (voiceless consonant spoken in a syllabic CV context) or (ii) at sound extinction (HBB consonantal sound). The TextGrid annotations were imported into MATLAB, in order to determine the analysis windows of the acoustic signal. The EMA and RIP signals were imported and processed in MATLAB. Spatial trajectories of 7 coils placed on three flesh points of the tongue (apex/blade - TA, middle - TM, and dorsum - TB) and four flesh points of the lips (upper and lower, median and lateral) were extracted from the EMA recordings and their speed was computed. The analyses were restrained to the mid-sagittal plane, therefore lateral tongue coils TR and TL were disregarded. The mean trajectories and their covariances were computed on the different sound repetitions. Median and interquartile range were extracted from breathing data of the /putikati/ and **PtKt** items. Time was normalized to account for differences among repetitions.

Regarding the statistical modeling, the impact of a 6-modality explanatory variable (p, t, k, kick, hi-hat, rimshot) on the variation in intensity or duration was tested. A linear mixed model was developed using the function *lme* of the package *nlme* of the statistical software R. Such a mixed model allows accounting for, simultaneously, the repetition of the measurements, the inter-beatboxer variability, and the residual variance which can change

from one consonant to another. For the duration variables (absolute duration and relative duration of the consonantal sound) whose distribution is skewed, their logarithm was used. A contrast analysis was conducted on the established models with the *glht* function of the *multcomp* package, following the method presented by Hothorn et al. (2008). The aim of the multiple comparisons is to explore whether there is a difference between a spoken consonant and its HBB equivalent.

5.3 Preliminary observations, or beatboxing speech

5.3.1 Phonetic and acoustic remarks

Table 5.3 summarizes a phonetic overview of the target sentences as they were expected and how they were actually realized in speech and HBB.

Table 5.3: Target sentences and their spoken and HBB realizations.

Target	Speech	HBB	IPA
boots and cats (/ˈbutsɛnˈkatsɛn/)	[ˈbutsɛnˈkatsɛn]	P^m t K t	[p'^m t' k' t']
des petits cookies des gros cookies (/depətikuˈkidegʁokuˈki/)	[deptikuˈkidegʁokuˈki]	P t K t t P K t	[p' t' k' t' t' p' k' t']
pâtes au pesto (/ˈpatope ˈsto/)	[ˈpato ˈpesto]	P t Pf t	[p' t' p'͡f t']

The spoken sentences were very close to the expected realizations. Minor adaptations were observed in the rhythm and in the syllabic structure. In the "Pâtes" sentence, the primary stress of the word *pesto* was expected on the last syllable (/peˈsto/), as per French phonetic stress rule (Walker, 1975). It was actually shifted on the second to last syllable ([ˈpesto]), resulting in a regular rhythm of one in two stressed syllables over the whole phrase. A similar rhythm was achieved in the "Cookies" sentence, which theoretically comprises 9 syllables. However, as is often the case in spoken French[3], the unstressed ə in the first syllable of the word "*petits*" (/pəˈti/) was deleted, and the two-syllable word was rearranged in a one-syllable word ([pti]). Hence, the sentence syllables were reduced to 8 (Fig. 5.1 middle right), with primary stresses on the fourth and eighth syllables. Conversely, a syllable ([ɛn]) was added at the end of the "Boots" sentence, making it a four-syllable phrase (Fig. 5.1 top right), with primary stress on the first and third syllable, as in the "Pâtes" sentence.

[3]At least in the variety of French spoken by PS.

The HBB rendition of the sentences was as follows:

1. "Boots": a sequence of post-voiced kick, closed hi-hat, rimshot, closed hi-hat;
2. "Cookies": a sequence of kick, closed hi-hat, rimshot, closed hi-hat, closed hi-hat, kick, rimshot, closed hi-hat;
3. "Pâtes": a sequence of kick, closed hi-hat, PF snare, closed hi-hat.

After the modifications of the spoken sentences discussed above, the number of syllables in the speech sentences and the number of boxemes in the HBB beats matched.

Upon visual inspection of the waveforms and spectrograms of the three spoken and beatboxed sentences (Fig. 5.1), it is apparent that little to no voicing is present in the HBB renditions. All vocalic sound has been suppressed and HBB consonantal sounds are separated by silence. Voicing is present as a post-effect of the kick in the "Boots" sentence, by way of a nasal bilabial occlusive. A formantic structure is also visible. Grossly, HBB sounds, especially rimshots, kicks, and snares, are characterized by higher levels of acoustic energy evenly distributed over all frequencies, whereas speech sounds show higher energy concentrations in the vowels at specific frequency bands (formants). Sibilants show higher energy concentration at higher frequencies.

Bilabials were produced in three different phonetic contexts ([bu], [pa], [pes]). As mentioned, voicing of the speech syllable [bu] was rendered in HBB by a nasal bilabial occlusive following the kick. A short silence is present between the kick and the nasal sound (Fig. 5.2 top), whereas in the speech syllable, consonant and vowel are continuous events. However, this is consistent with an ejective production of the kick, where the silence between the consonantal and nasal sounds would indicate the time necessary for the glottis to release the occlusion of the glottalic initiation mechanisms and for the vocal folds to eventually meet the conditions necessary for auto-oscillation. The CVC structure of the syllable [pes] (Fig. 5.2 bottom) is transformed in an affricate sound (PF snare) that can be decomposed into two segments. The first sounds similar to a plosive, but presents a marked vibratory aspect (see discussion in section 4), however the acoustic outcome is much different than that of a bilabial trill of speech. The second segment is clearly of turbulent, but not sibilant nature. Spectral intensities are much higher than those of the speech sounds. In general, these bilabial HBB kicks and snare have a strong burst (amplitude of waveform) and impulse attack and are longer than the speech consonants. The kick sounds more 'punchy', whereas the PF snare sounds more 'dry', likely due to a different center of gravity.

HBB alveolars (Fig. 5.3) are characterized by an ascending/descending transient noise, very different from the impulse attack (short burst) of the speech counterparts. The higher

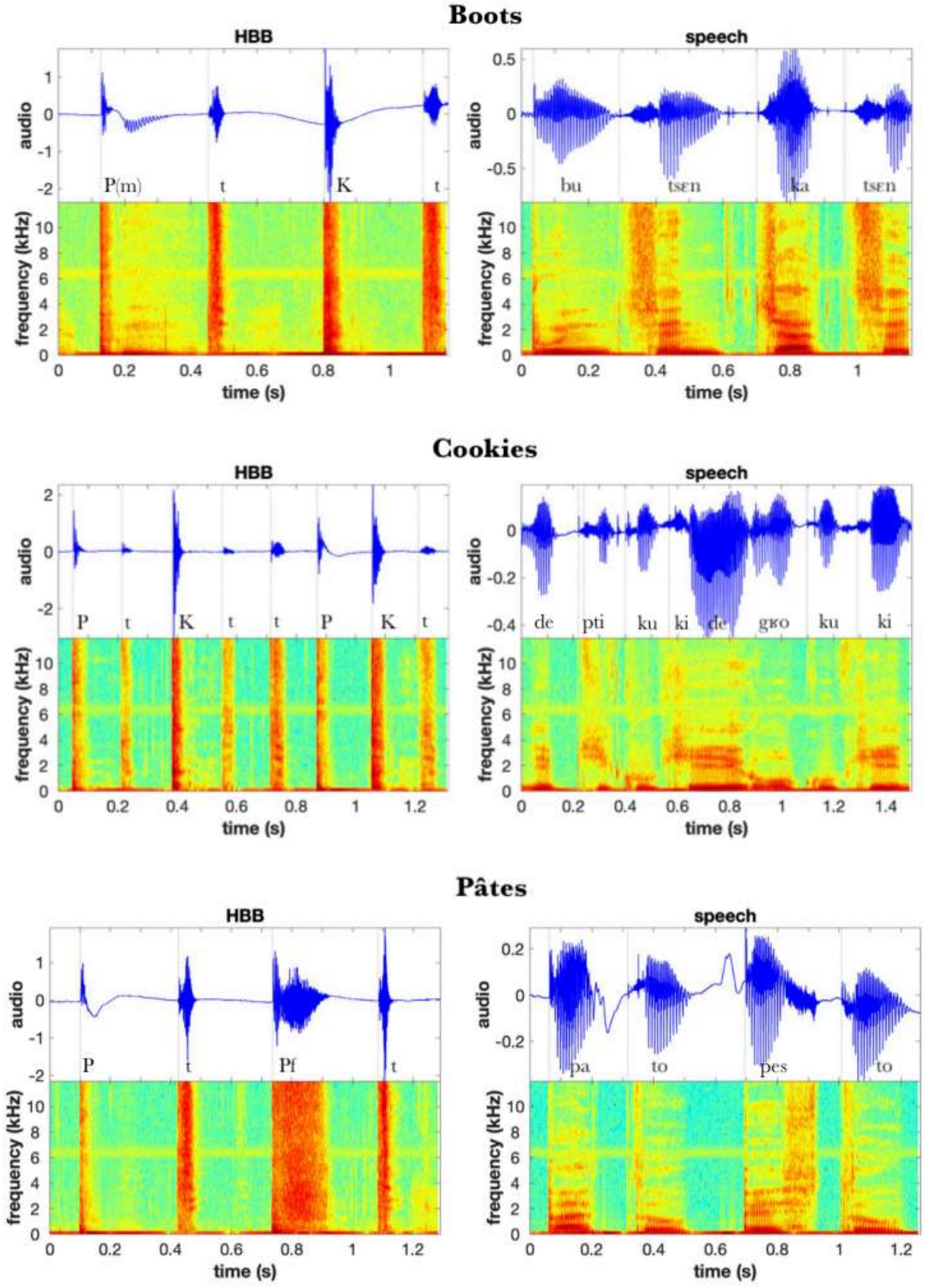


Figure 5.1: Audio waveforms and broadband spectrograms of a representative token of the three sentences produced as HBB (left) and speech (right). Spectrogram parameters: view-range: 0-12 kHz; window length: 5 ms; dynamic range: 50 dB.

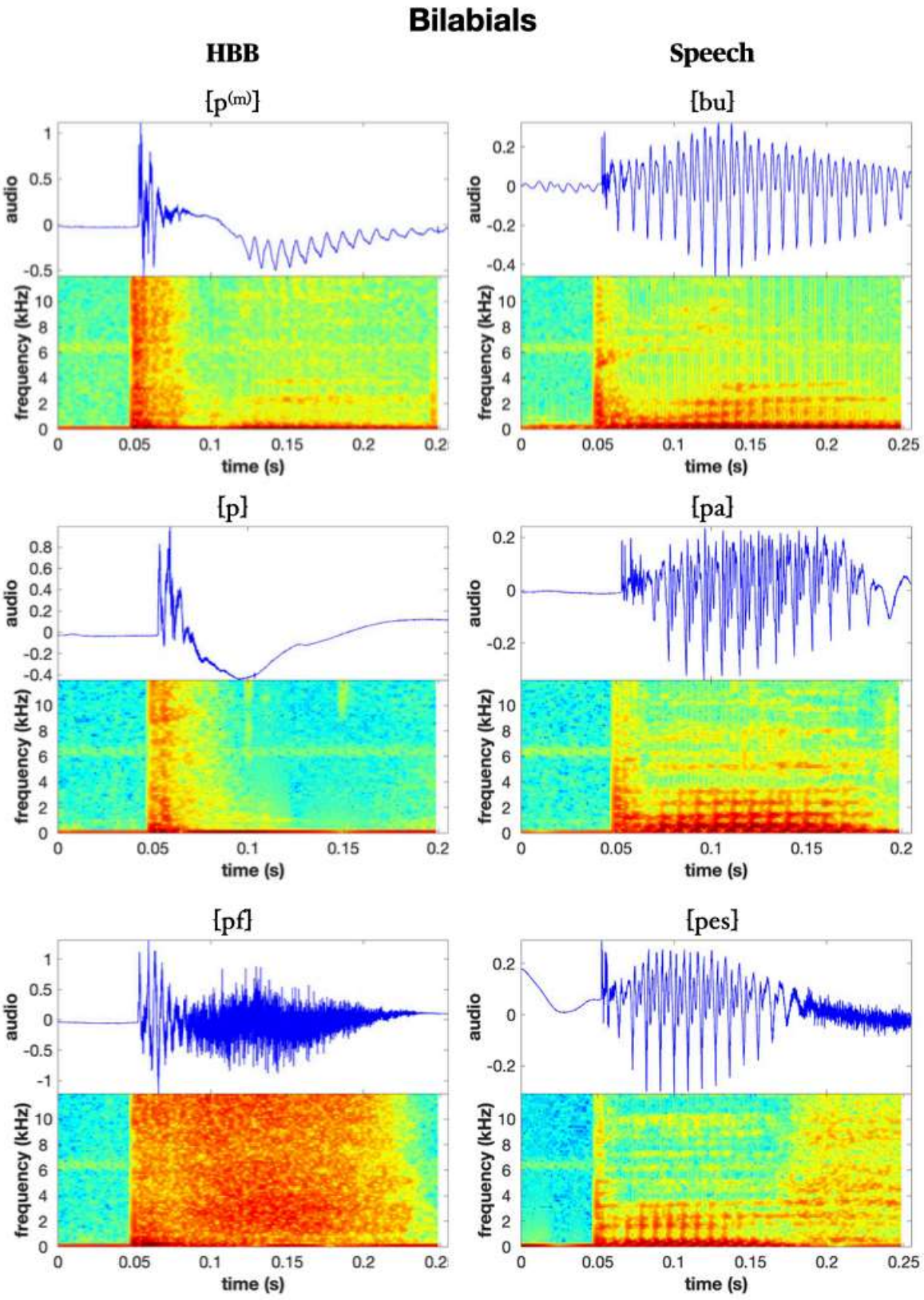


Figure 5.2: Audio waveforms and spectrograms of a representative token of the realization of bilabial sounds. Top: bilabials from the Boots sentence; middle: bilabials from the Cookies sentence; bottom: bilabials from the Pâtes sentence. Spectrogram parameters: view-range: 0-12 kHz; window length: 5 ms; dynamic range: 50 dB.

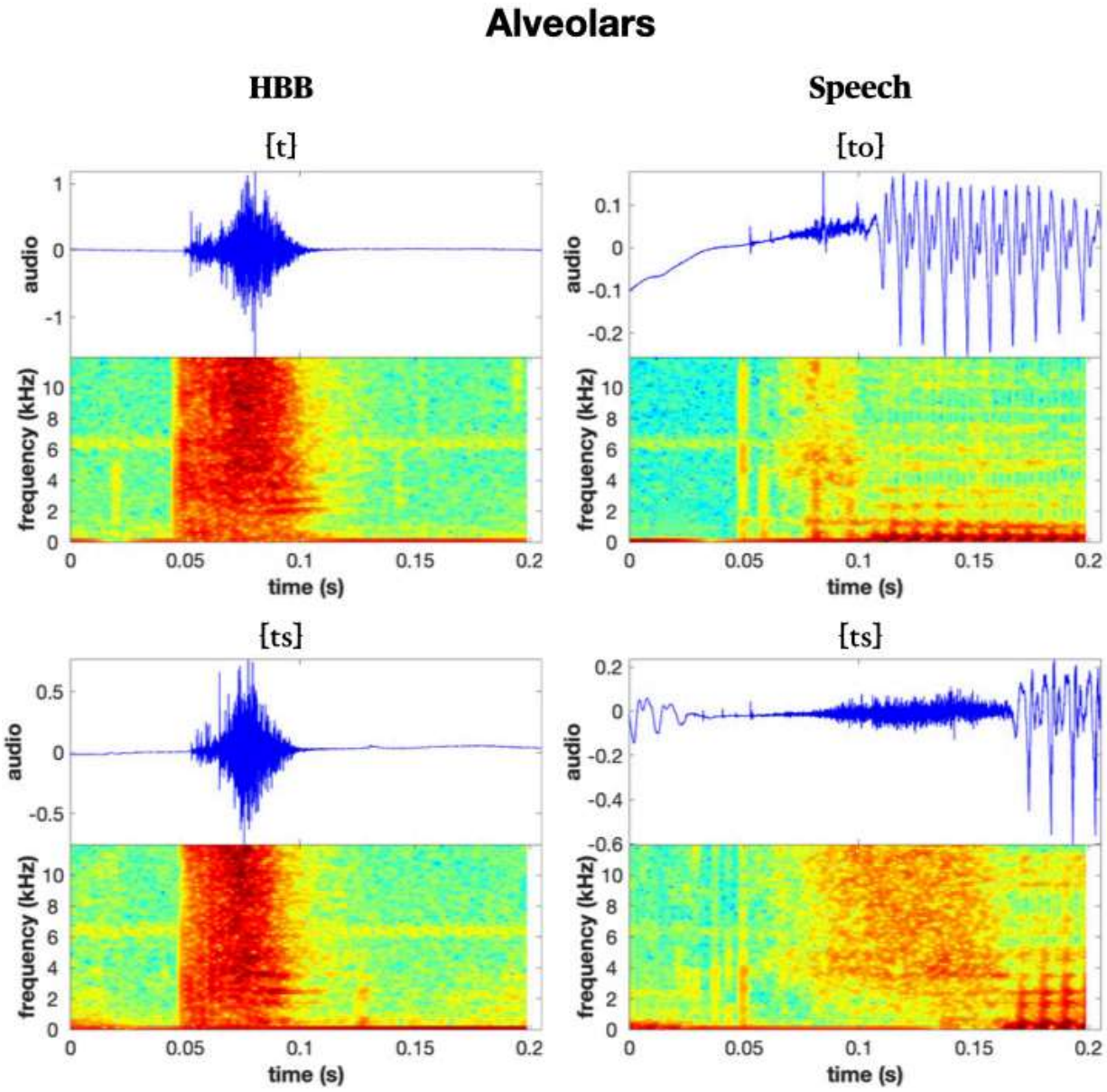


Figure 5.3: Audio waveforms and spectrograms of a representative token of the realization of alveolar sounds. Top: alveolars from the Pâtes sentence; bottom: alveolars from the Boots sentence. Spectrogram parameters: view-range: 0-12 kHz; window length: 5 ms; dynamic range: 50 dB.

amplitude in the waveform translates into higher acoustic energy, especially in the middle part of the sound. Spectral bands appear towards the end of the sound. The speech alveolars are characterized by a visible, but not so intense burst and by a protracted release noise whether a fricative segment is present ([ts]) or not ([to]). Moreover, the occlusion phase is not always silent. This can be caused by incomplete occlusion due to the presence of the TIP coil combined with low articulatory force, but this does not happen with HBB sounds.

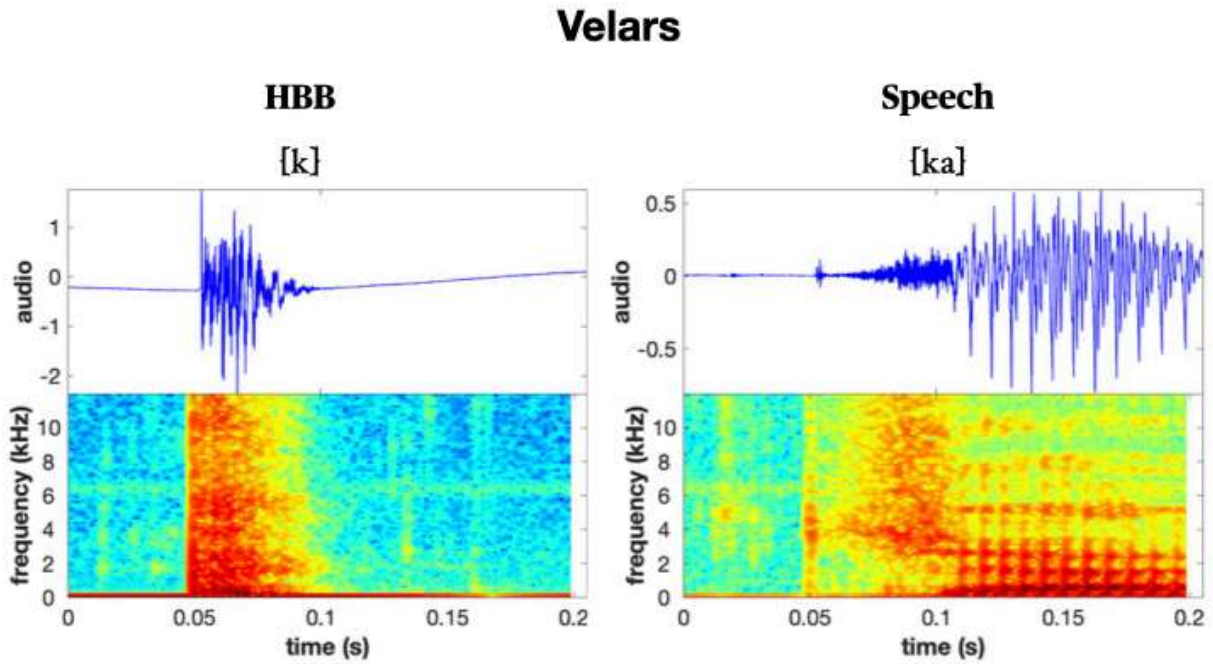


Figure 5.4: Audio waveforms and spectrograms of a representative token of the realization of velar sounds from the Boots sentence. Spectrogram parameters: view-range: 0-12 kHz; window length: 5 ms; dynamic range: 50 dB.

HBB velars (Fig. 5.4) have a very strong and clean burst, with more energy in the lower frequencies and spectral bands towards the end of the sound. They have the largest waveform amplitude of all the sounds. Speech velars are characterized by a less strong burst and a long release noise, that becomes more intense towards the vowel.

In sum, the acoustic signature of HBB sounds has different characteristics from that of the corresponding speech consonants. In contrast with speech where consonants and vowels are alternated, in HBB only consonantal sounds are present. The presence of vocal fold vibration is very limited, but other sources of vibration (the lips) may be introduced. Lastly, HBB sounds have more acoustic energy over a larger scale of frequencies than speech consonants.

5.3.2 Articulatory behavior

The observations on the articulatory kinematics hereby presented are based on the analysis of the evolution in time of the trajectory of a given coil moving in space, as well as its speed. This analysis is visualized over a time window of 400 ms for the sentence "Cookies" and 800 ms for "Boots", centered on the burst. The analysis presented here is restrained to the vertical (y) axis indicating the low-high movements.

The preliminary investigation of PS data (EMA and video) shows that the place of articulation of each of the plosives remains roughly the same in speech and in HBB.

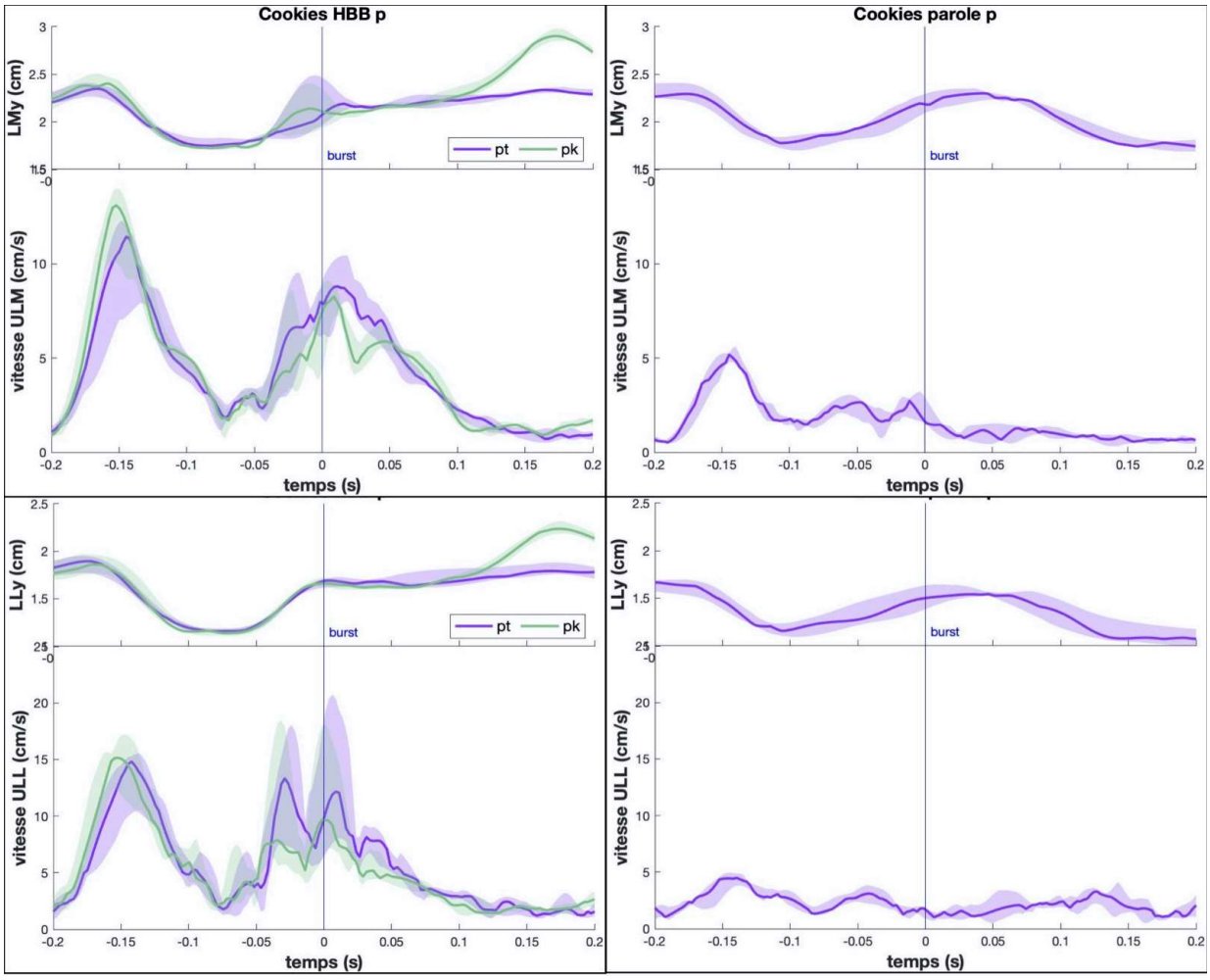


Figure 5.5: Top: median and interquartile range of the mid-lip (LM), left-lip (LL) intercoil distance along the y-axis, computed from all occurrences of the bilabial plosive for the sentence Cookies and its beatboxed counterpart. Bottom: median and interquartile range of the tangential velocity (3D) of the coil of interest.

Figures 5.5 and 5.7 illustrate the kinematics underlying the realization of the plosive /p/ and its beatboxed equivalent, kick, in the "Cookies" sentence in its spoken version "des **p**'tits cookies ..." or in the HBB beat "**P** t K t t **P** K t". Both sounds are obtained by complete occlusion of the lips, attested by visual inspection of the video data. In both production modes (spoken and beatboxed), the occlusion forms about 100 ms before the burst, as attested by the minimum of interlabial distance (LMy) in Fig. 5.5. In speech, a reproducible articulatory behavior is observed for the different repetitions of the sentence. It is highlighted on Figure 5.5 by a low interquartile range. A more variable behavior is observed in HBB near the burst for this same place of articulation (middle of the lips). It is interesting to note that a homogeneous behavior (low interquartile range) similar to the one observed in speech is found on the left half of the lips (LL coils). This suggests that the release of this bilabial drum sound is lateralized to the left, which is confirmed by the video data (Fig. 5.6).

Figure 5.6: Occlusion release of the bilabial sounds.

Further, the interlabial distance (LLy) is held at a minimum over 50 ms before the lips begin to part (rise in LLy). This does not occur for the spoken plosive, where the minimum interlabial distance is briefly reached, but not held. This might suggest a more secure occlusion in HBB, where the lips are more tightly pressed together. The release speed of the bilabial occlusion is clearly greater in HBB than in speech, with an increase in the speed of the coil attached to the middle of the upper lip (ULM on Fig. 5.5) which can reach 10 cm/s around the burst, as opposed to 2.5 cm/s in speech. As for the left upper lip coil (ULL on Fig. 5.5), closer to the place of occlusion release in HBB, it can reach 20 cm/s at the time of the burst, while in speech its speed is almost zero.

A difference between speech and HBB is also observed in the realm of lingual kinematics, as illustrated on Figure 5.7 for the coil placed on the back of the tongue (DORS). If the movement of this coil follows the burst in speech, it precedes it in HBB. A marked vertical displacement of the back of the tongue occurs in HBB, accompanied by a very low variability of the trajectory at the time of the burst, but different depending on the sound that follows, as a result of coarticulation.

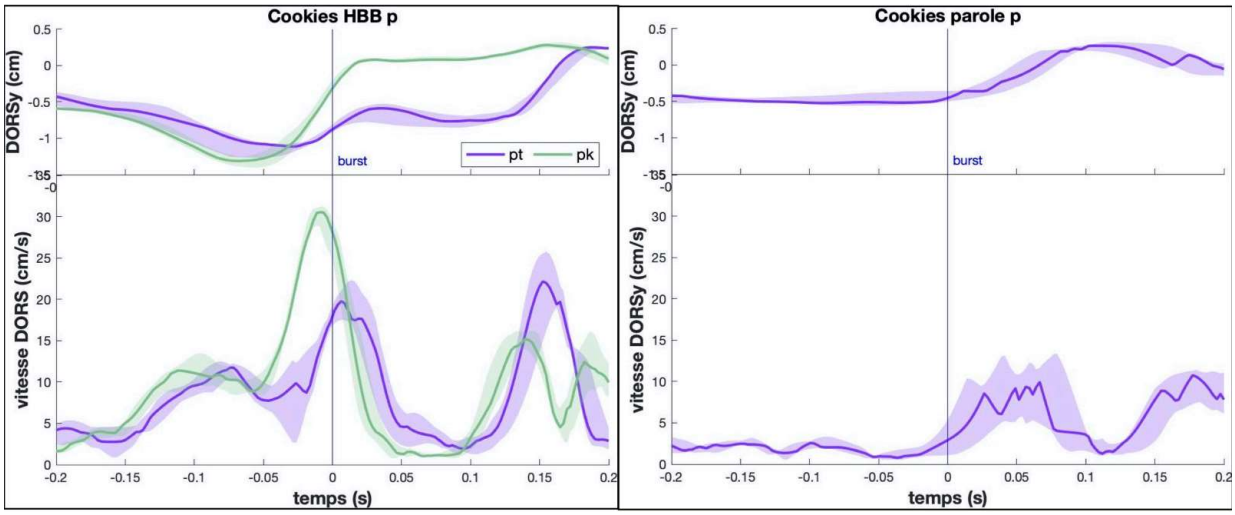


Figure 5.7: Top: median and interquartile range of the DORS coil along the y-axis, computed from all occurrences of the bilabial plosive for the sentence Cookies and its beatboxed counterpart. Bottom: median and interquartile range of the tangential velocity (3D) of the DORS coil.

Figure 5.8 illustrates the articulatory kinematics of the tongue tip for the production of the plosive /t/ and its HBB equivalent, hi-hat, in the case of the same sentence "p **t**its cookies ..." and the beat "P **t** K **t t** P K **t**". A well-marked plateau is observed in HBB in correspondence of which the coil speed is almost zero. This suggests that the occlusion of the CV takes place in the alveolar or post-alveolar region and is maintained over roughly 100 ms. This plateau is not found in speech. In both production modes, the TIP coil commences its downward motion before the burst, detected in the audio signal. In both cases, the variability of the trajectory is reduced. In particular, the interquartile range is small before the burst. Although less pronounced than in the kick case, the speed of movement of the coil of interest at release is greater in HBB than in speech.

Concerning the plosive /k/ and its rimshot equivalent in HBB, the case is presented of the "Boots" sentence in its spoken version "boots and **c**ats", or beatboxed as follows: "$P^{(m)}$ ts **K** ts".

In speech as in HBB (Fig. 5.9), the trajectories of the back of the tongue (DORS coil)

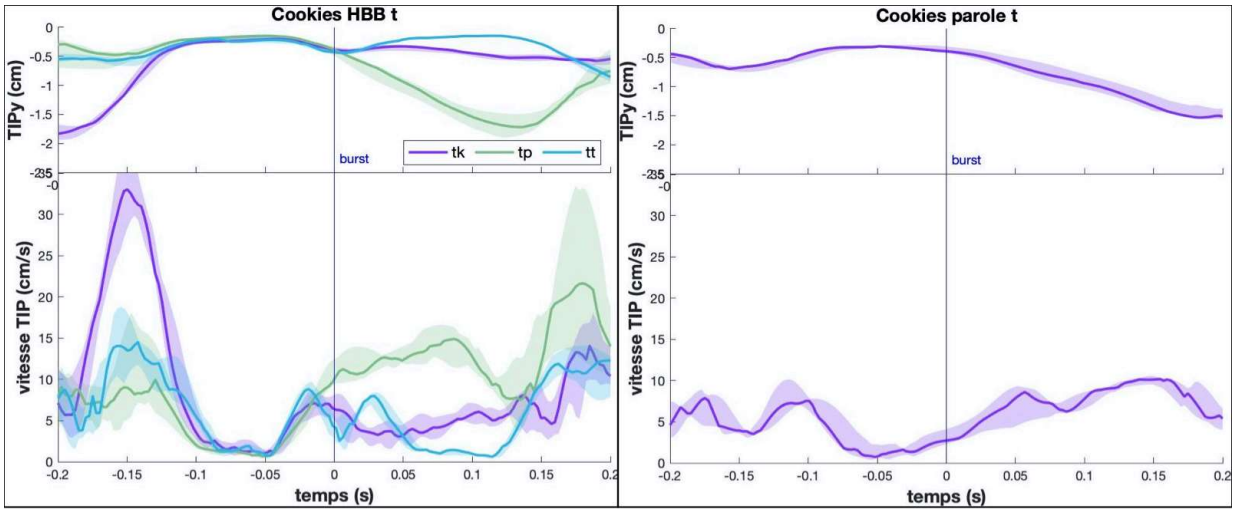


Figure 5.8: Top: median and interquartile range of the TIP coil along the y-axis, computed from all occurrences of the alveolar plosive for the phrase Cookies and its beatboxed counterpart. Bottom: median and interquartile range of the tangential velocity (3D) of the TIP coil.

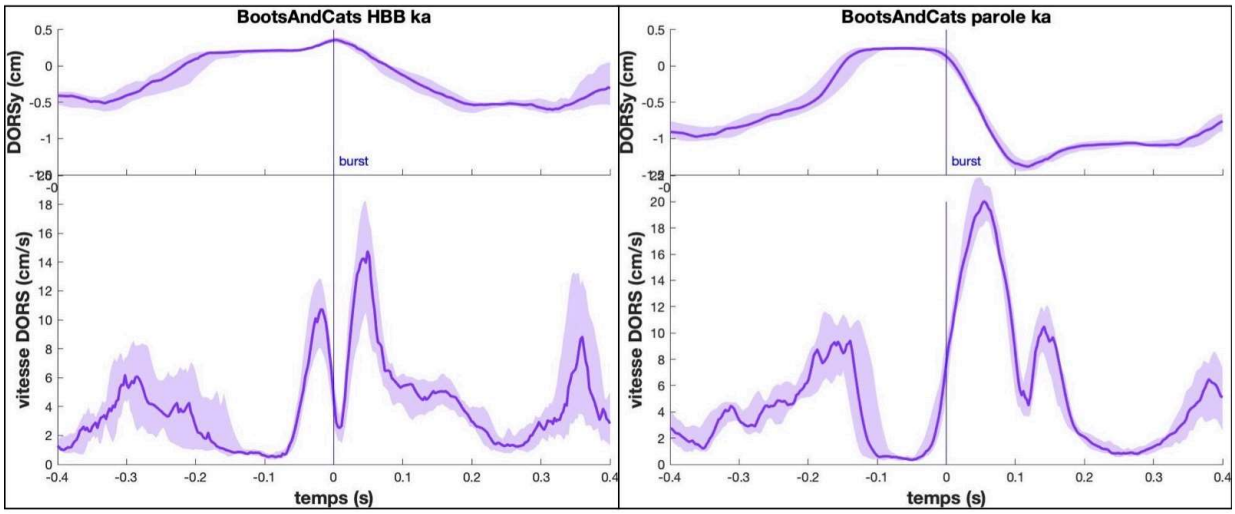


Figure 5.9: Top: median and interquartile range of the DORS coil along the y-axis, computed from all occurrences of the velar plosive for the sentence BootsAndCats and its beatboxed counterpart. Bottom: median and interquartile range of the tangential velocity (3D) of the DORS coil.

show a plateau before the burst, indicating the occlusion of the CV at the palato-velar region. In speech, this plateau is maintained over 60 ms until the burst, when the coil shows a downward movement toward the next vowel position ([a]). In HBB, the plateau is almost 100 ms long. Further, a supplementary movement of the DORS coil is observed before the burst: towards the end of the occlusion, the coil undergoes an upward movement that ends in correspondence with the burst. This movement is not insignificant, as indicated by the very small interquartile range. It induces a velocity peak before the burst (about 10 cm/s), which is not observed in speech. Indeed, in speech, the only velocity peak around the burst is the one related to the displacement of the tongue from the velar occlusion to the much lower position of the open front vowel [a]. In HBB, two velocity peaks are observed, one just before and the other just after the burst, which means that at the time of the burst the DORS coil is actually decelerating.

These preliminary observations show that, even though it is possible to glimpse a common root, especially at the level of the place of articulation and the general trajectory, the production of percussion sounds in HBB is clearly distinct from that of plosive sounds in speech.

The articulation details of HBB drum sounds reveal an articulatory kinematics that is unique to HBB and different from that of speech. The shapes of coil trajectories in the region of articulation during the occlusion phase seem to suggest that HBB sounds have longer and/or more stable occlusions, which might be explained by more secure occlusions, where the articulators are more tightly pressed together to endure higher aerodynamic pressure behind the occlusion. Further, in the case of the kick compared to a voiceless bilabial plosive, the occlusion release is lateralized. This lateralization of the release is observed in some beatboxers, but not systematically (see sec. 5.4.1). The type of laterality (left or right) depends on the beatboxer. This labial gesture is probably intended to better control the tension of the lips at the moment of release. The tongue is also very active even if the occlusion, i.e., the main place of articulation, occurs at the lips. The speed of movement of the articulators is often greater in HBB than in speech, especially before the burst.

The lingual kinematics of the HBB percussion sounds appear to be consistent with the use of an ejective or glottalic initiation mechanism. The highly active and reproducible lingual dynamics during articulation of the HBB kick sound do not seem to be completely explained by coarticulation, since with equal phonetic environment in speech (especially /pt/), no comparable movements are observed. Moreover, all three coils of the tongue are involved in almost the same way, indicating a global movement of the tongue. This seems rather to suggest a movement of the tongue in relation to the laryngeal rise proper to a glottalic initiation. Indeed, several studies have attested to the use of this mechanism in the articulation of drum sounds in HBB (Blaylock et al., 2017; De Torcy et al., 2014; Dehais Underdown et al., 2019; Dehais-Underdown et al., 2021; Patil et al., 2017; Proctor

et al., 2013; Sapthavee et al., 2014).

The same is true for the upward movements of the tongue at the end of the occlusion phase of the velar HBB percussion sounds. As for the nature of these lingual movements, Proctor and colleagues (2013) hypothesize that the tongue is used in concurrence with the larynx to produce a more efficient pushing action. Indeed, an optimized glottalic initiation mechanism could increase the acoustic efficiency of the sound produced, which is very important in HBB.

On the other hand, these preliminary data did not show this type of movement during the articulation of the alveolar plosives of any of the three beatboxed sentences. This could be explained by the fact that in these cases the rise of the larynx has no influence on the tongue because of its articulatory position (apex higher than the back). In general and to our knowledge, few studies have explored the impact of the ejective mechanism on the lingual dynamics, in speech as in HBB.

By means of a very active articulatory kinematics and the wide use of this mechanism, the study of the production of HBB percussion sounds makes it possible to highlight articulatory phenomena at the lingual level.

5.3.2.1 Space exploration

The differences in articulatory behavior highlighted so far for each sound raise the question of how each coil, and therefore each related point on the tongue surface, moves in space during HBB production as compared to speech production. If the production mechanisms are different for HBB and speech sounds, does that mean that the oral articulators move differently inside the oral cavity? Do they explore the same or different oral regions? Do the lips and the jaw movements happen in the same spatial regions? Given the observations of different articulatory behaviors and extra lingual movements for HBB as compared to speech, it seems reasonable to formulate the hypothesis that the coils may move in different regions inside and outside the oral cavity depending on whether the mode of production is HBB or speech. In order to verify and further specify this hypothesis, all the trajectories of the 8 coils were considered during the production of the 3 sentences "Boots", "Cookies", and "Pâtes". First, density charts were produced for HBB and speech production regrouping articulatory data from each sentence, i.e. three density charts for HBB and three for speech. However, the repetitions per sentence were too few, and therefore the resulting maps were not informative. Subsequently, data from the three sentences were grouped into one desnity chart for HBB and one for speech. Figure 5.10 shows the density chart concerning "Boots", "Cookies", and "Pâtes" produced in HBB (48 total repetitions) and in speech (32 total repetitions).

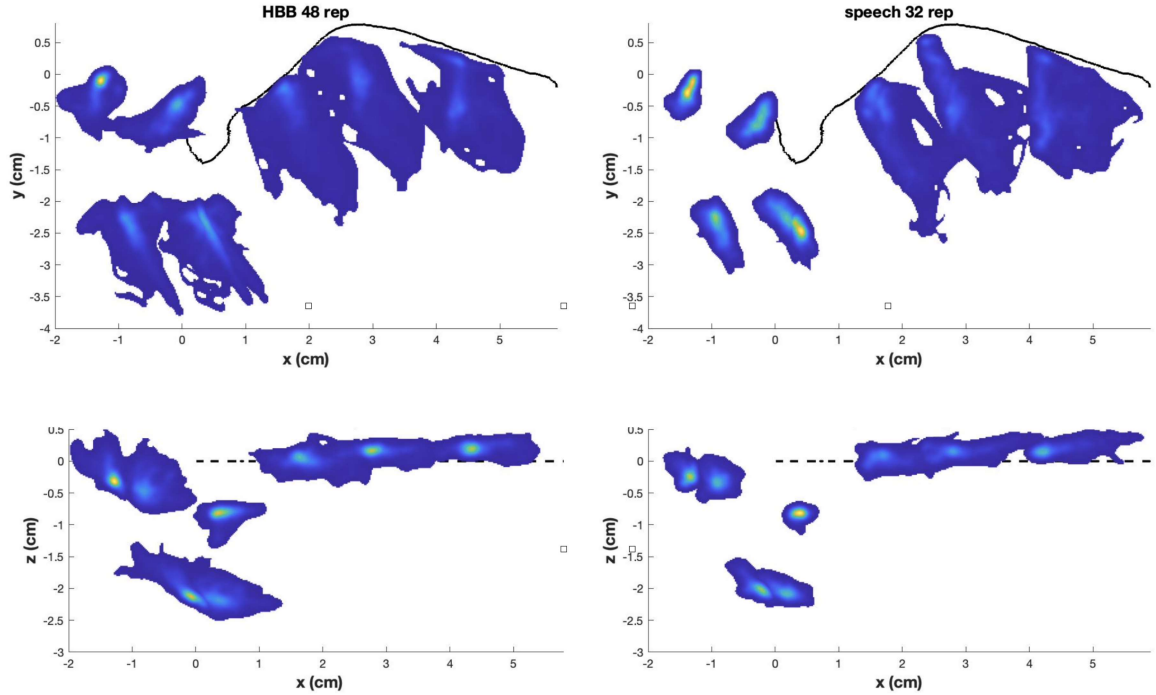


Figure 5.10: Density charts of HBB (left) and speech (right) production. Top row: sagittal (xy) view; bottom row: transversal (zx) view. Solid black line: palate contour. Left: front; right: back.

It is evident that the exploration of the intraoral and perioral space is different in HBB and in speech. In speech, the movements appear more stereotyped (higher presence frequencies, in yellow), especially in the perioral region (lips). In HBB, the exploration of a larger space is noticeable, however the frequency of passage in a given point (or pixel) is lower (darker blue). Concerning HBB, the coils of the tongue explore larger regions near the palate, especially towards the front. In addition to vertical movements, forward movements of the tongue are also observed, that might be interpreted as related to the upward movement of the larynx during an ejective articulation mechanism. The coils of the lips explore larger regions, vertically, but also and especially transversally (x-axis). This might be interpreted as the effect of possibly higher air flows that produce more visible aerodynamic drive on the lips. Of course, aerodynamic measures are needed to confirm or infirm this. Regarding speech production, the tongue coils show more stereotyped and linear movements. These movements appear to be diagonal, following the anatomy of the tongue. The lip coils show more stereotyped movements as well, with no evidence of major aerodynamic effects.

The superposition of the two density charts (Fig. 5.11) reveals that in HBB the tongue is generally in a higher position and explores a space that is higher than that explored in

speech production. As for the lips, they explore a more lateral space.

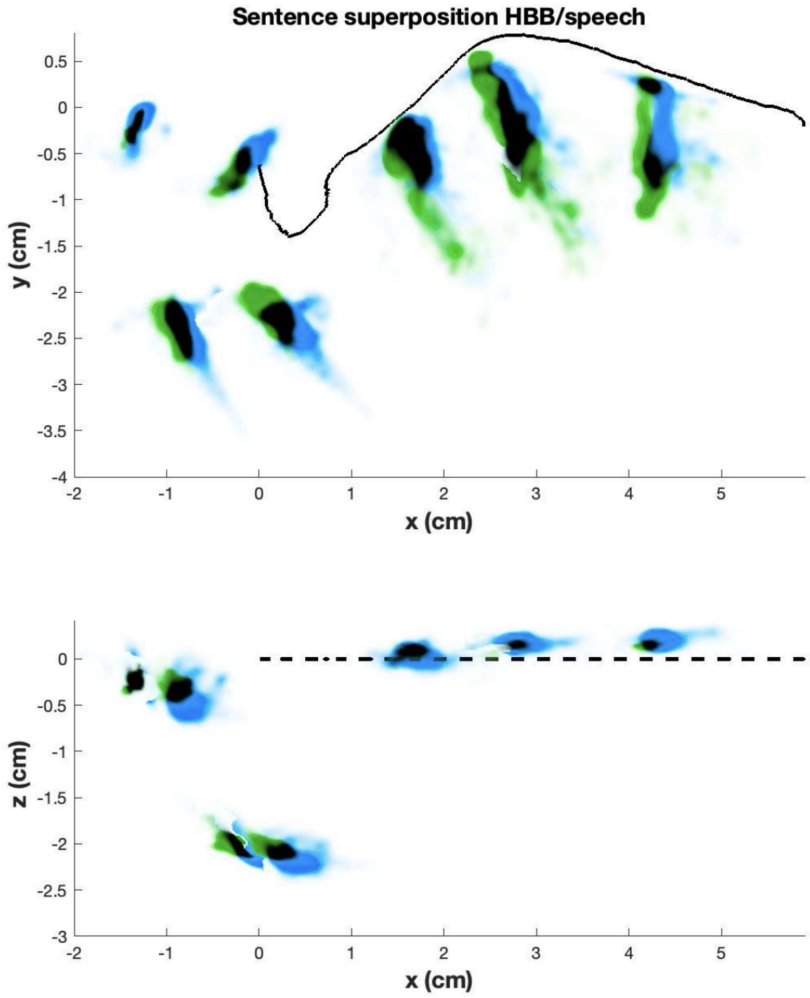


Figure 5.11: Superposition of HBB (blue) and speech (green) density charts. Top row: sagittal (xy) view; bottom row: transversal (zx) view. Solid black line: palate contour. Left: front; right: back.

5.3.3 Breathing behavior

Figure 5.12 shows a global view of RV variation over time during the whole task of sentence repetition[4].

[4]Data regarding the transition phase go beyond the scope of the work hereby presented, and therefore are not discussed.

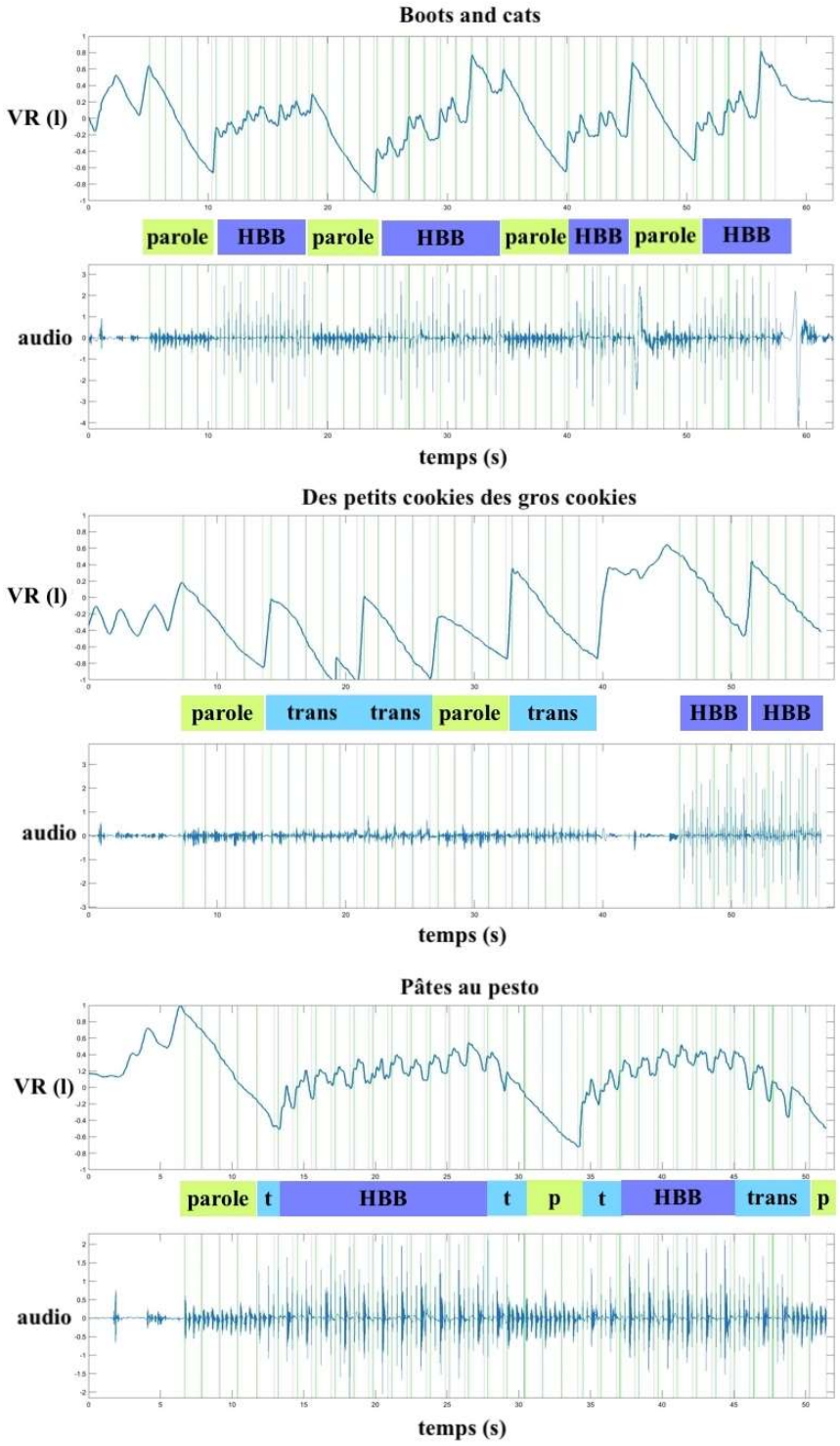


Figure 5.12: Evolution over time (in seconds) of reconstituted volumes (VR, in liters) and corresponding audio signal. Vertical lines indicate the beginning (green) and the end (black) of one sentence repetition.

It is apparent that at least two different respiratory behaviors are employed at different times: during spoken production, RV rapidly increases before the beginning of sound production (inhalation) and it slowly and steadily decreases during sound production (exhalation); as for the second type of behavior, the respiratory signal undergoes slight local variations throughout sound production, but sometimes these variations can be large and RV can increase very rapidly. In general, the first behavior seems to be characteristic of spoken productions and the second of HBB.

Focusing on single breath cycles (BC), intended as RV between two major inhalations, the two different behaviors appear even more clearly (Fig. 5.13). Spoken sentences are repeated 4 ("Boots") to 5 ("Pâtes") times, over an exhalation phase that can last 6 ("Boots") to 7 ("Pâtes") seconds. Breathing cycles involving HBB sentences can sustain vocal production up to 12 repetitions ("Pâtes") over 16 seconds. In speech, the breathing cycle starts with a more or less deep inhalation that rapidly increases the RV from values around 0 to more positive or even strongly positive values; the sound production starts in correspondence with the RV peak and continues uninterrupted while the RV decreases steadily until it reaches strongly negative values. In HBB, the breathing cycle also starts with an inhalation that makes the RV rise from strongly negative values to values slightly below 0, but still negative. Throughout the cycle, local variations of RV are observed in correspondence with the sounds produced. In general, in HBB the RV tends to vary slightly around an average value close to 0 or to even increase as in the case of BC2 (Fig. 5.13). Regardless, during HBB sound production the RV values never reach strongly negative values and rather tend to increase over the breathing cycle. It is noticeable how the breathing strategy can change within the same HBB breathing cycle, unlikely what happens in speech. For instance, repetition 4 and 6 of BC2, 2 and 4 of BC3, and 2 and 4 of BC4 (Fig. 5.13) show a stabilisation of RV in the middle part of the sentence, that may be interpreted as a vocal production in apnea. Further, rapid and deep increases in RV can appear towards the end of a sentence, but before the end of vocal production. This indicates that, in these instances, the last boxeme (namely a hi-hat) is produced via an ingressive airstream. The amplitude of RV variation ($\sim$0.8 l) suggests that the airstream is likely pulmonic.

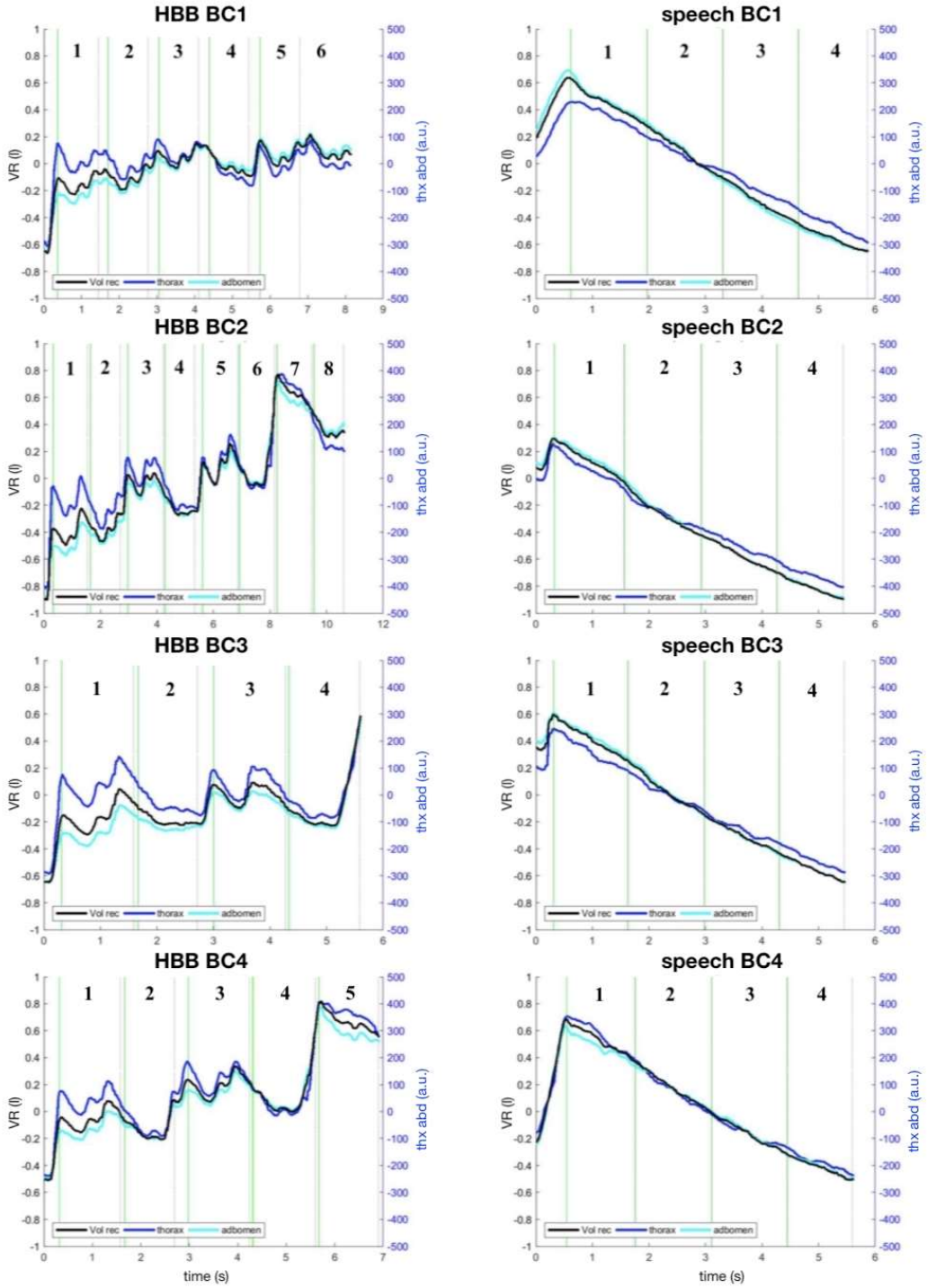


Figure 5.13: Evolution over time of respiratory volumes related to *Boots and cats* per breath cycle (BC). HBB is displayed on the left, speech on the right. Reconstituted volumes (VR) are expressed in liters, thorax and abdomen signals in arbitrary units, time in seconds. Vertical lines signal the acoustic start (green) and end (black) of a sentence repetition. Numbers indicate the progression of sentence repetition.

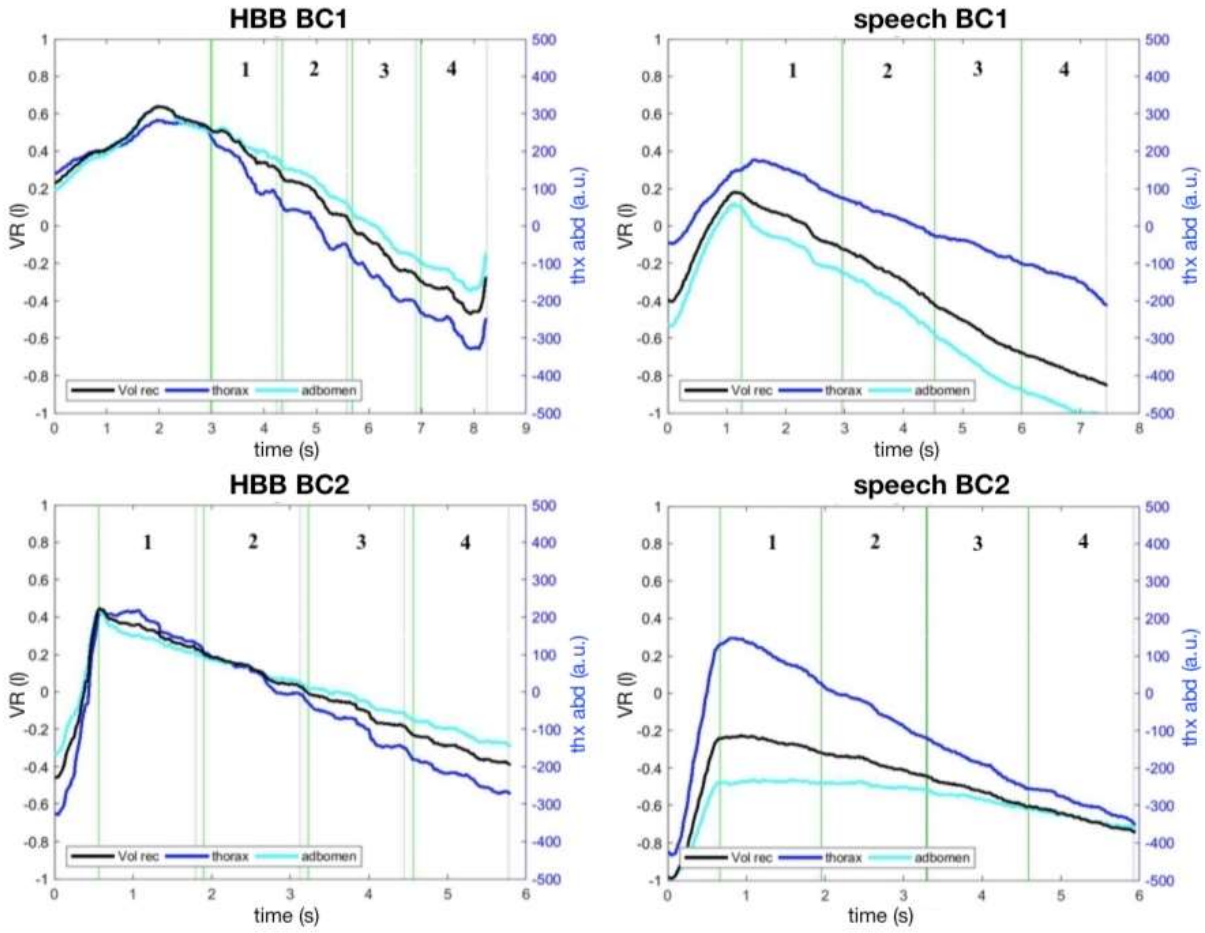


Figure 5.14: Evolution over time of respiratory volumes related to *Cookies* per breath cycle (BC). HBB is displayed on the left, speech on the right. Reconstituted volumes (VR) are expressed in liters, thorax and abdomen signals in arbitrary units, time in seconds. Vertical lines signal the acoustic start (green) and end (black) of a sentence repetition. Numbers indicate the progression of sentence repetition.

Data relative to "Cookies" sentence show a different scenario from that described above: the respiratory behaviors show a similar general trend between HBB and speech (Fig. 5.14). Concerning speech, the behavior is very comparable to the one observed for the other sentences: a rapid increase of the RV (about 0.6-0.8 liters) before the beginning of sound production and a regular decrease throughout sound production. In contrast to what was observed for the other sentences, the respiratory behavior of HBB closely resembles that of speech, with a large increase in RV before the onset of sound production and a more or less regular decrease, but nonetheless less regular than speech, during sound production. However, the contributions to the RV of the thoracic and abdominal compartments are dissimilar. Both the compartments are roughly equally engaged in HBB production,

whereas in speech the majority of the contribution comes from the thoracic compartment. Nevertheless, local RV variations are present in HBB production.

Figure 5.15 quantifies the variation of RV over each repetition of the sentence with respect to the RV at the onset of the repetition ($\Delta RV(t) = RV(t) - RV(0)$). Time is normalized to account for differences in the duration of various repetitions. Speech shows the most reproducible behavior among tasks, with a quasi linear volume variation. The higher volume values are always in correspondence with the beginning of the phonation at the beginning of the sentence. The volume variations occur almost without exception below 0, in the negatives, which means that the volume gradually decreases during the sentence.

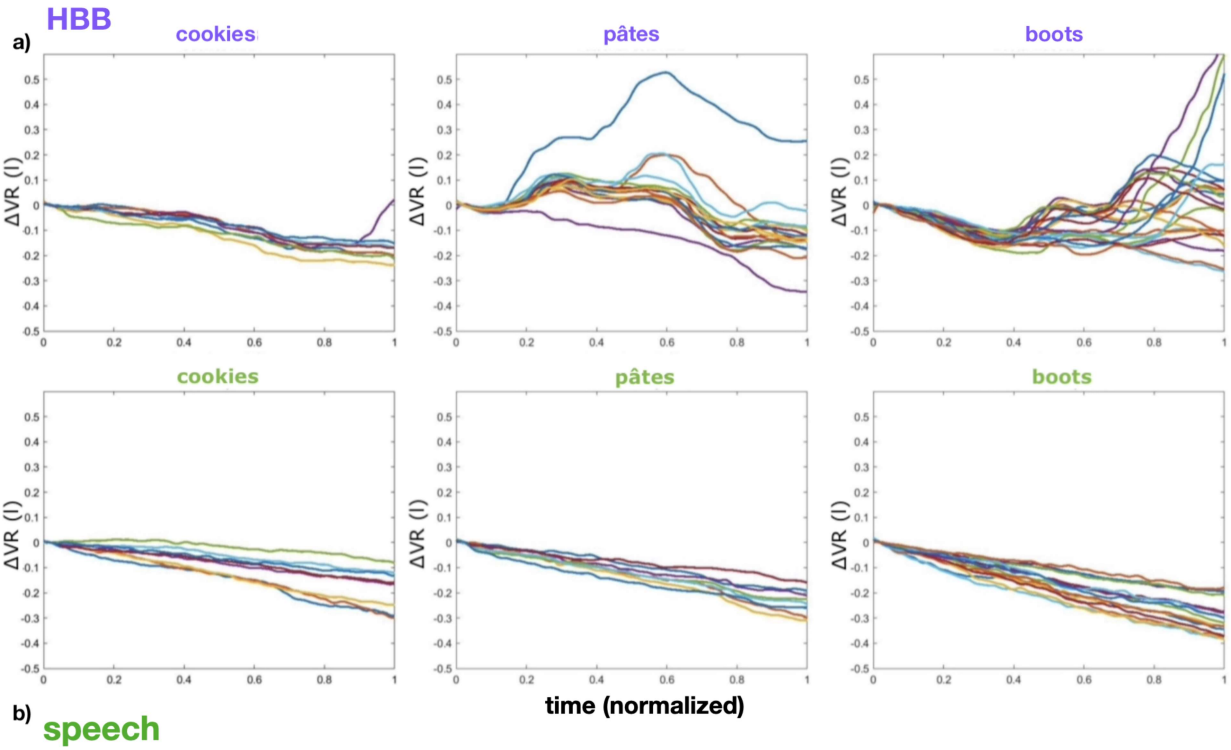


Figure 5.15: Variation of respiratory volumes (ΔVR, in liters) over a sentence repetition. HBB is displayed on top, speech bottom. Time is normalized to account for differences in duration of repetition.

In contrast, in the case of HBB, the variation of respiratory volume during a sentence is hardly linear or uniform. As previously mentioned, the most similar behavior to speech is that observed for the "Cookies" sentence. In this instance, however, the RV decreases more rapidly between 0.4 and 0.7 time units. The RV variations at the end of the sentence are negative in all but one case and range from -0.15 to -0.25 l. Therefore, RV at the

end of the sentence is generally lower than at the beginning of the sentence. Nevertheless, for the other two sentences, positive variations of VR are observed. Regarding "Boots", during the first part of the sentence up to 0.4 time units, VR decreases ($\Delta VR < 0$). Then, three different behaviors can be identified: a first behavior where the RV does not increase anymore until the end of the sentence; a second one where the RV increases towards the end of the sentence sometimes up to very high values and more often up to values as high as those at the beginning of the sentence; and a third behavior where RV increases a first time towards the middle of the sentence and again towards 0.8 time units, to then decrease towards the end of the sentence. As opposed to speech, where ΔVR is always negative and between -0.1 and 0.4 l, in HBB ΔVR can be either negative or positive and vary between a larger range (-0.25 – 0.6 l). As for the "Pâtes" sentence, in HBB, after a slight decrease of the ΔRV curves at the beginning of the sentence, indicating a decrease in RV, and therefore an egressive airstream, ΔRV becomes positive between 0.1 - 0.2 and 0.6 time units, indicating an increase in RV with respect to the initial RV. After 0.6 time units, a significant decrease of the curves occurs, up to 0.8 time units, indicating an egressive airstream, and then a small plateau until the end of the sentence. Two curves clearly show an atypical behavior, one above the others, indicating higher RV variations, but temporally compatible with the variations of most curves, the other below, showing a different behavior compared to the other curves. In general, the final values of the RV variation curves are in the negatives, between values just below 0 and -0.2 l, showing that the final RV values are smaller compared to the initial values.

Central tendencies reflect these observations (Fig. 5.16). Spoken sentences show a quasi-linear decrease in RV over the repetition with limited variability (low interquartile range). Similar is the case for "Cookies" produced in HBB mode, with even less variability. On the contrary, a higher degree of variability (bigger interquartile range) characterizes the other two HBB sentences, more so "Boots". This, of course, is due to the use of different airstream mechanisms (namely, ingressive or egressive) from one repetition to another. Further, in the HBB "Pâtes" sentence, the tendency is that of an increase in RV in the middle of the sentence as compared to the RV at the beginning of the sentence. In the HBB "Boots" sentence, this increase in evident in the second half of the sentence, and the central tendency is that of finishing the sentence with a similar RV as the beginning.

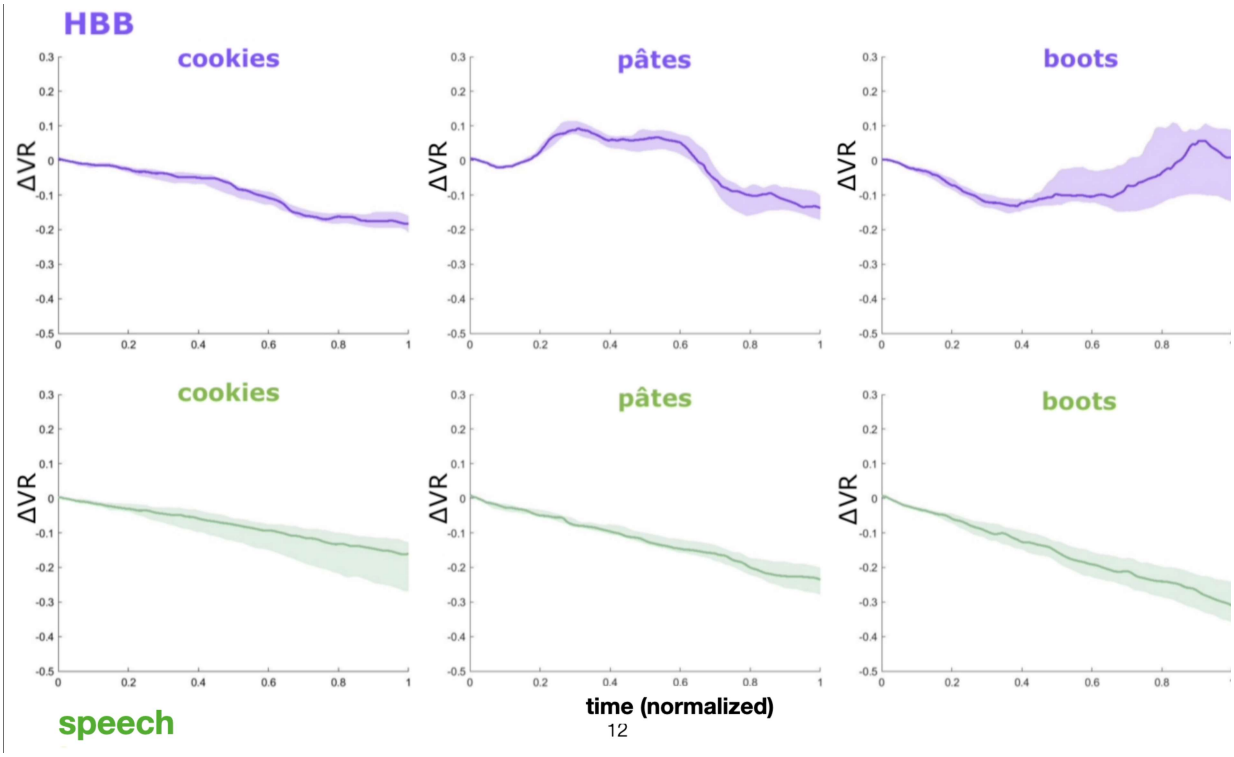


Figure 5.16: Median (solid line) and interquartile range (colored area) of variation of respiratory volumes (ΔVR, in liters) over a sentence repetition. HBB is displayed on top, speech bottom. Time is normalized to account for differences in duration of repetition.

To summarize, the results show that PS uses a characteristic respiratory strategy for HBB, that is different from that associated with speech. The breathing behavior during spoken tasks has similar characteristics to those described in the literature: a large increase in RV before the onset of phonation, indicating an inhalation, followed by a gradual, quasi-linear and generally uninterrupted decrease during phonation, corresponding to an exhalation. This results in a highly asymmetric RV curve during a breathing cycle. Thus, the breathing cycle consists of two phases, one, faster (in the order of a second), the inhalation, and the other, more extended in time (5-7 seconds), the exhalation that sustains phonation. This behavior means that phonation ends at lower values of RV than those at the onset of phonation, and this with low variability. On the other hand, HBB is associated with a non-linear evolution of RV, characterized by numerous positive and negative variations throughout the breathing cycle, most of which are small in amplitude. These variations reflect the presence of micro-inhalations throughout the breathing cycle and a true continuous alternation of inhalations and exhalations likely aimed at maintaining RV within a relatively small range of variation around volume values that are neither too high nor too low. However, strongly positive RV variations can occur during phonation, especially towards the borders of the breathing cycle. This increases variability in breathing

behavior, since the same boxeme can be produced by means of an egressive or ingressive airstream. This respiratory strategy makes a deep inhalation before the beginning of phonation not necessary and in fact phonation can begin without any prior inhalation. In addition, if a deeper-than-usual inhalation occurs at the beginning of the breathing cycle, it never raises the RV to high values. Further, RV at the end of the breathing cycle may be greater than that at the onset of phonation, indicating a net volume gain during phonation. This seems to be consistent with Tiko's suggestion of the possibility of "too much air" in HBB [5]. In any case, such an alternation of inhalations and exhalations makes the notion of breathing cycle as used so far not at all adapted with such a breathing behavior. Indeed, micro inspirations scattered throughout the sound production and a maintenance of RV above strongly negative values prevent the beatboxer from running out of breath and bypass the need to interrupt phonation to perform a deeper inhalation. As a result, phonation can be protracted over long periods of time. Of course, such an alternation of inhalations and exhalations would not be suitable for speech where consonants and vowels are coarticulated, producing a mostly continuous acoustic signal. In contrast, in HBB, at least in the case of the three sentences tackled in the present work, no vocalic sound is present, and, acoustically, the consonantal sounds are separated from each other by brief silences of the order of a few hundred ms.

5.4 More evidence from more beatboxers

5.4.1 Acoustics and articulatory behavior

Figure 5.17 demonstrates the acoustic and spectral signatures of the spoken syllables and their boxeme counterparts. The third token of the repetitions is chosen as representative.

Boxemes are systematically associated with an acoustic wave of higher amplitude than spoken consonants. This difference was quantified by calculating the intensity of the consonantal sound. Results are visualized in Fig. 5.18. The intensity difference is strongly significant for each pair of sounds: overall and on average, [p] is softer than **P** by 20.9±1.6 dB ($p<0.001$), [t] is softer than **t** by 12.7±1.4 dB ($p<0.001$), [k] is softer than **K** by 16.0±1.4 dB ($p<0.001$).

Figure 5.19 shows the measures of the mean duration and their standard deviations for the four subjects and the three comparisons.

The difference is significant for all the pairs of sounds: bilabial (-0.6134±0.096, $p<0.001$), apico-alveolar (-0.2522±0.0978, $p<0.05$), and velar (-0.6782±0.0936, $p<0.001$).

[5] Private conversation, January 2018.

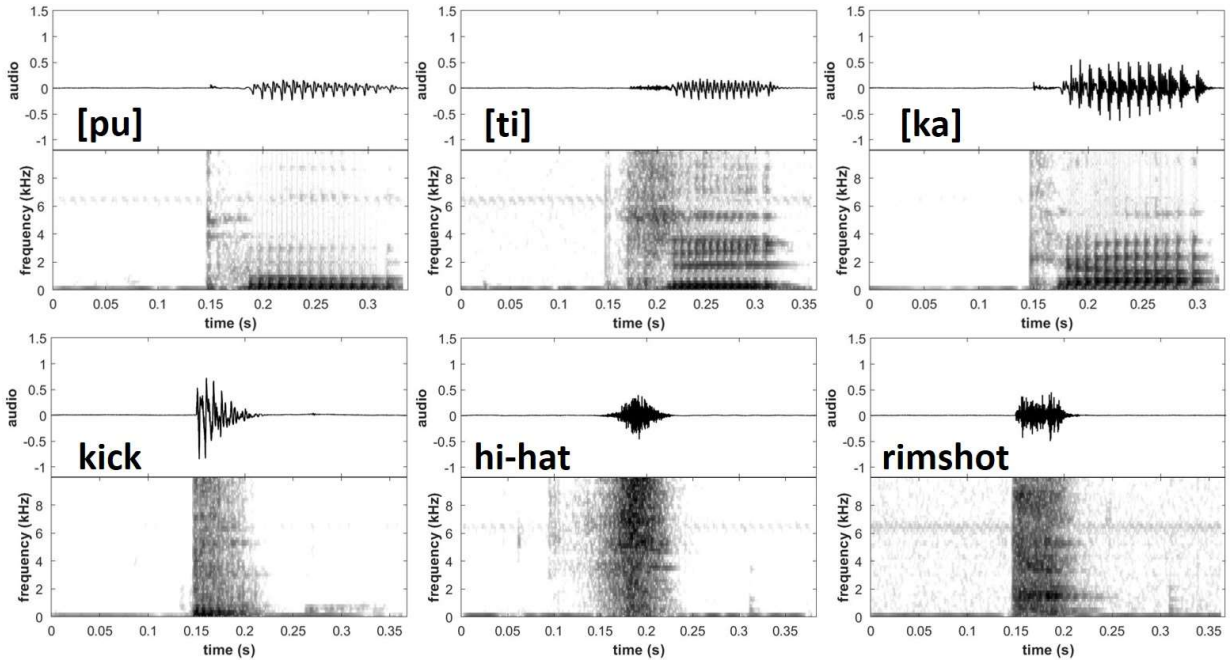


Figure 5.17: Audio waveforms and spectrograms of a representative token (3rd repetition) of each boxeme and corresponding consonant of S04. Spectrogram parameters: view range: 0-10 kHz; window length: 9 ms; dybamic range: 30 dB.

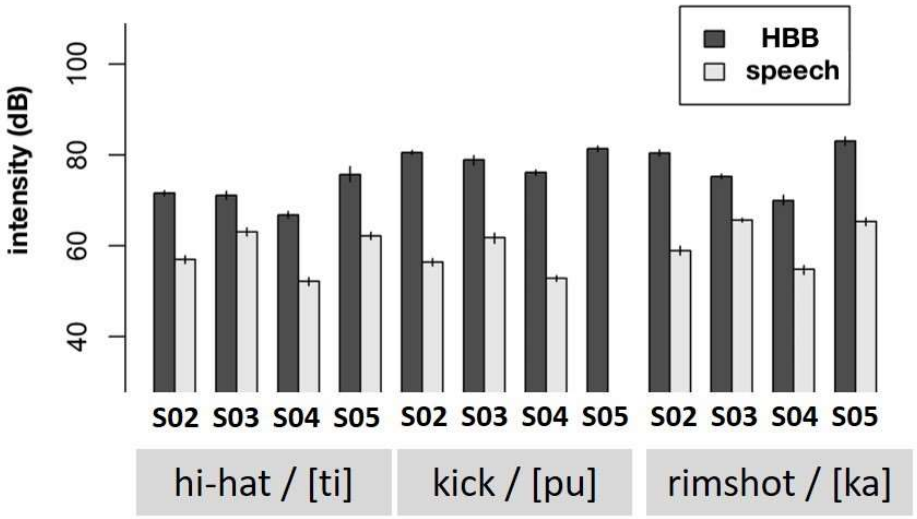


Figure 5.18: Mean and standard deviation of acoustic intensity (in dB) of the three boxemes and corresponding consonants produced by the four beatboxers (S02-S05).

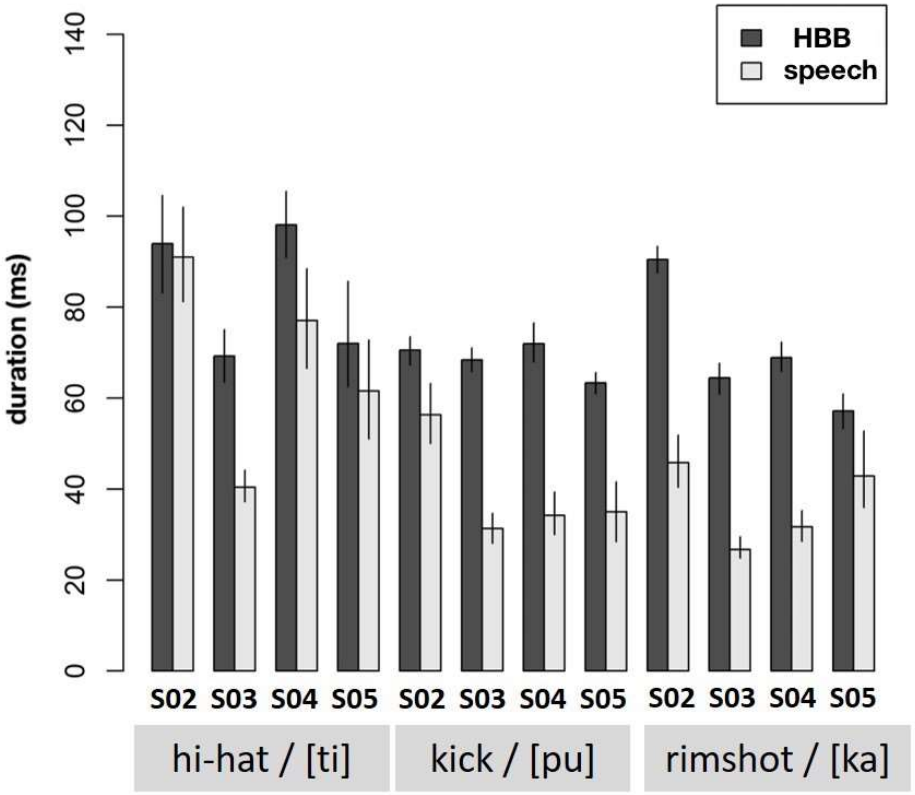


Figure 5.19: Mean duration (in ms) of boxemes and consonants for the four subjects.

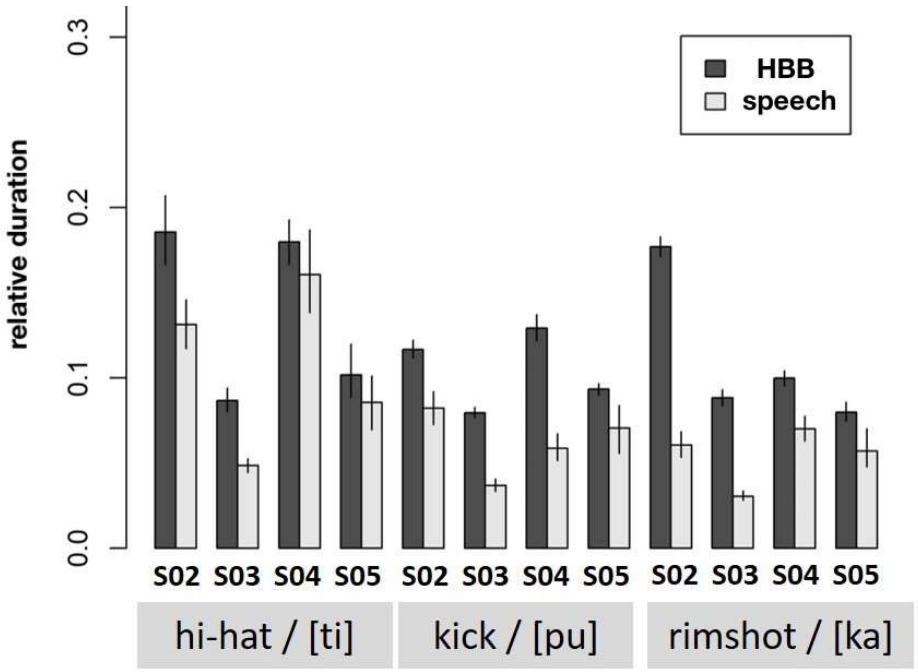


Figure 5.20: Duration of boxemes and consonants relative to the tempo for the four beatboxers.

Because the rhythm of the consonantal repetition was left to the beatboxer's discretion during the sequence, the duration of the consonantal sound was also related to the duration of a repetition cycle, thus taking into account the natural tempo of the sound repetition sequence. The differences are even more pronounced for these duration ratios (Fig. 5.20). As before, the difference is significant for all pairs of sounds: bilabial (-0.5758±0.1259, $p<0.05$), apico-alveolar (-0.3182±0.1274, $p<0.001$) and velar (-0.7283±0.1243, $p<0.001$).

The inspection of the articulatory data confirmed that the plosives of the spoken syllables and their beatboxed counterparts shared a similar place of articulation. The kick sound was produced as a bilabial plosive, such as the [p] in /pu/. The hi-hat was produced as a alveolar, in the same articulatory region as [t]. The rimshot was produced in the dorsal region, similar to [k]. Figure 5.21 illustrates the case of S03.

However, the EMA technique does not allow to determine the precise point of the occlusion. It will therefore not necessarily be exactly the same between beatbox and speech. For instance, in general, in the case of **K** the TB coil comes into contact with the palate further forward than in the case of [k] (see Fig. 5.21).

The spoken bilabial plosive was released centrally by all the beatboxers (**Figure 5.22**). In contrast, not all the beatboxers released the kick sound on the mid-sagittal plane. S01 systematically released the bilabial occlusion of the kick laterally, on the right side.

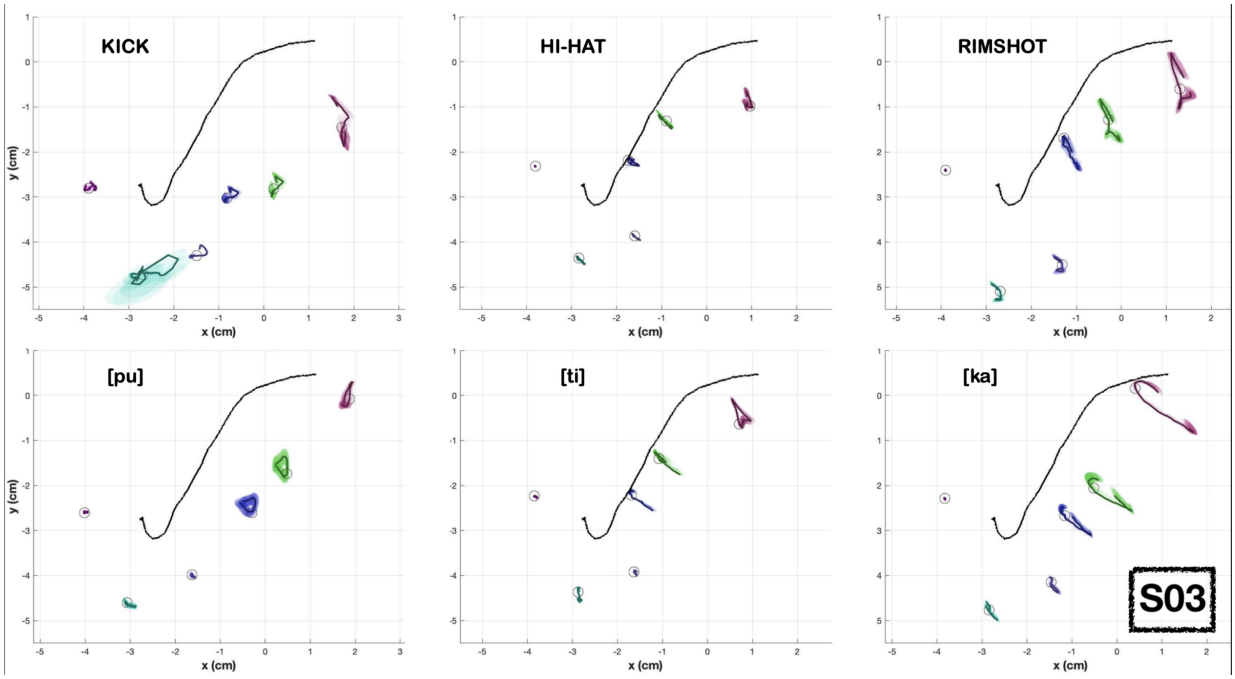


Figure 5.21: Illustration of the mean articulatory trajectories involved in the production of beatboxed (top) and spoken (bottom) consonantal sounds and their variance for subject S03. Circles indicate the time of acoustic burst. Visualized time window: 300 ms before and after the burst.

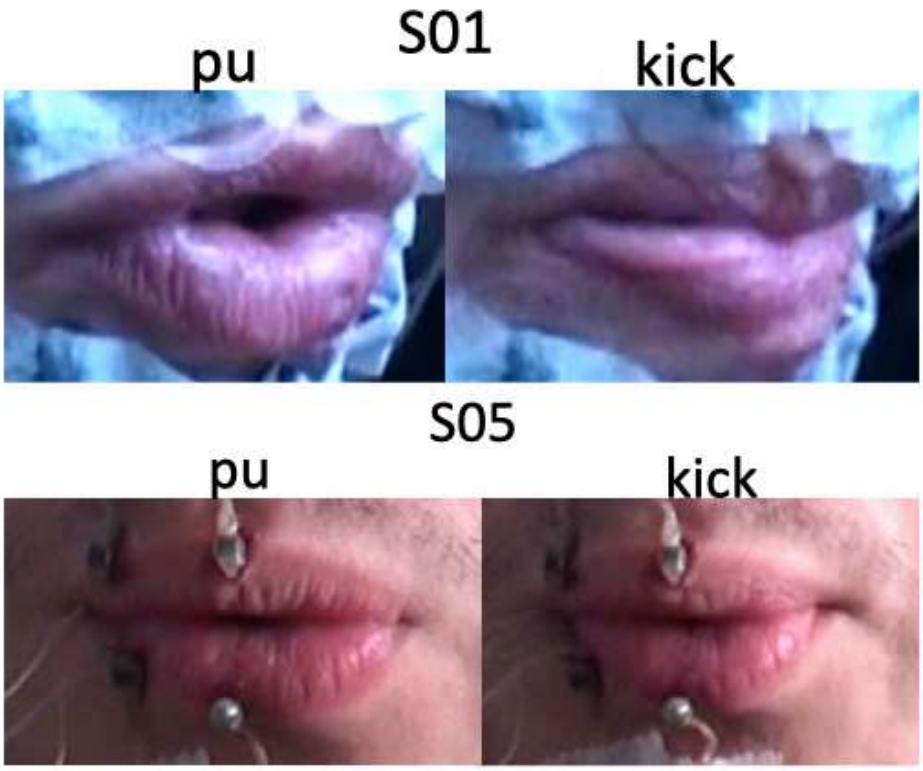


Figure 5.22: *Release of the bilabial sounds.*

In general, HBB sounds were produced with more ample and rapid movements than speech sounds, notwithstanding the absence of coarticulation with vocalic sounds. This was particularly the case in the [pu] vs kick comparison.

The case of **P** is particularly interesting. During the acoustic production phase, the interlabial distance (visualized in Figure 5.23 by the LM plot, the vertical distance between the central coils of the upper and lower lips) varies rapidly and greatly after the release of the occlusion, whereas in the case of [p] the two lips never part considerably, even at the moment of the occlusion release. This very different lip movement is highlighted in Figure 5.21 for the medial coil of the lower lip.

As for the lips, the tongue shows fast movements, especially in the back region (TB coil). This is the case for all beatboxers, with TB reaching up to 20 cm/s, while the speeds measured for [pu] are around 5–6 cm/s. These upward movements start before the burst and continue until the end, or even after the sound has ended (Fig. 5.24 bottom). However, the tongue adopts a generally lower position inside the oral cavity compared to [p]. By superimposing the trajectories of all the occurrences of the sequence, the coils of the tongue draw loops related to these upward movements. Such loops are more marked in some beatboxers (S01, S03) than in others (S02, S04). During the articulation of [pu],

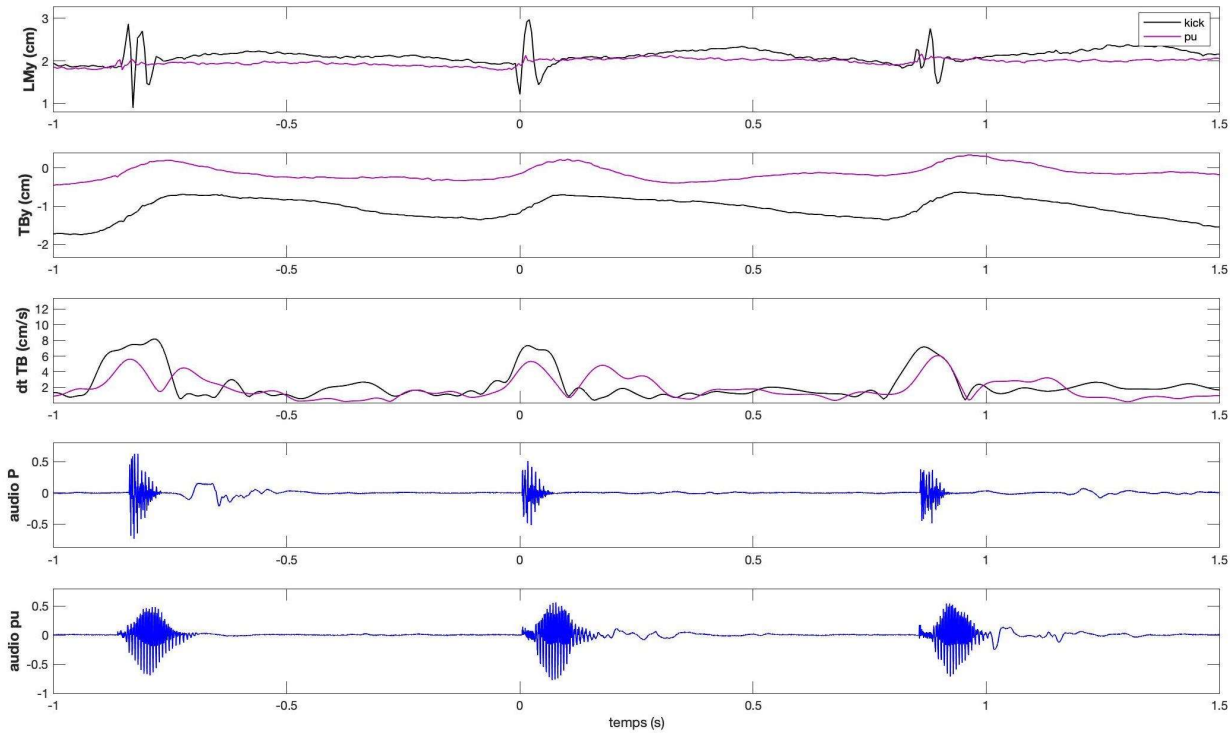


Figure 5.23: Spatial trajectories along the y-axis of the distance between the two central coils on the lips (LM), the tongue back (TB) coil, its time derivative (dt TB), and the audio signal of 3 repetitions of kick sounds and of [pu]. Black: HBB; purple: speech.

these loops can appear, but are often less pronounced and less stereotyped.

K showed the most varied articulatory strategies and most distant from that of speech. S02 and S04 created the occlusion in the region of TB, but much further back for S04 compared to S02, and the acoustic signature of the sound is very different. In S03 and S05, TA is in contact with the palate and remains so at the time of the release of the occlusion in the TB region. The contact can be broken (S03) between two subsequent articulations of the boxeme or may be kept in place (S05) during the whole sequence. The acoustic signature of the two sounds is quite different. Regardless of the different articulatory strategies, the articulatory speeds associated with **K** are systematically lower than those of [ka]. Articulatory loops are clearly visible in all beatboxers for the articulation of [ka]. They may appear in the case of **K** (S02-S04), but are generally less ample (S04) and are oriented along a more vertical axis than in speech (Fig. 5.21). A more similar articulatory behavior between HBB and speech is found in the case of **t**. The occlusion occurs in the TA region across subjects. However, in HBB, the tongue has a lower overall position and the TB coil may rise slightly after the burst. As for the articulatory speeds, two cases of figure are observed. In most cases (S02, S04, S05), the release at the level of TA is faster in HBB than in speech. However, in S03, the release is faster in speech.

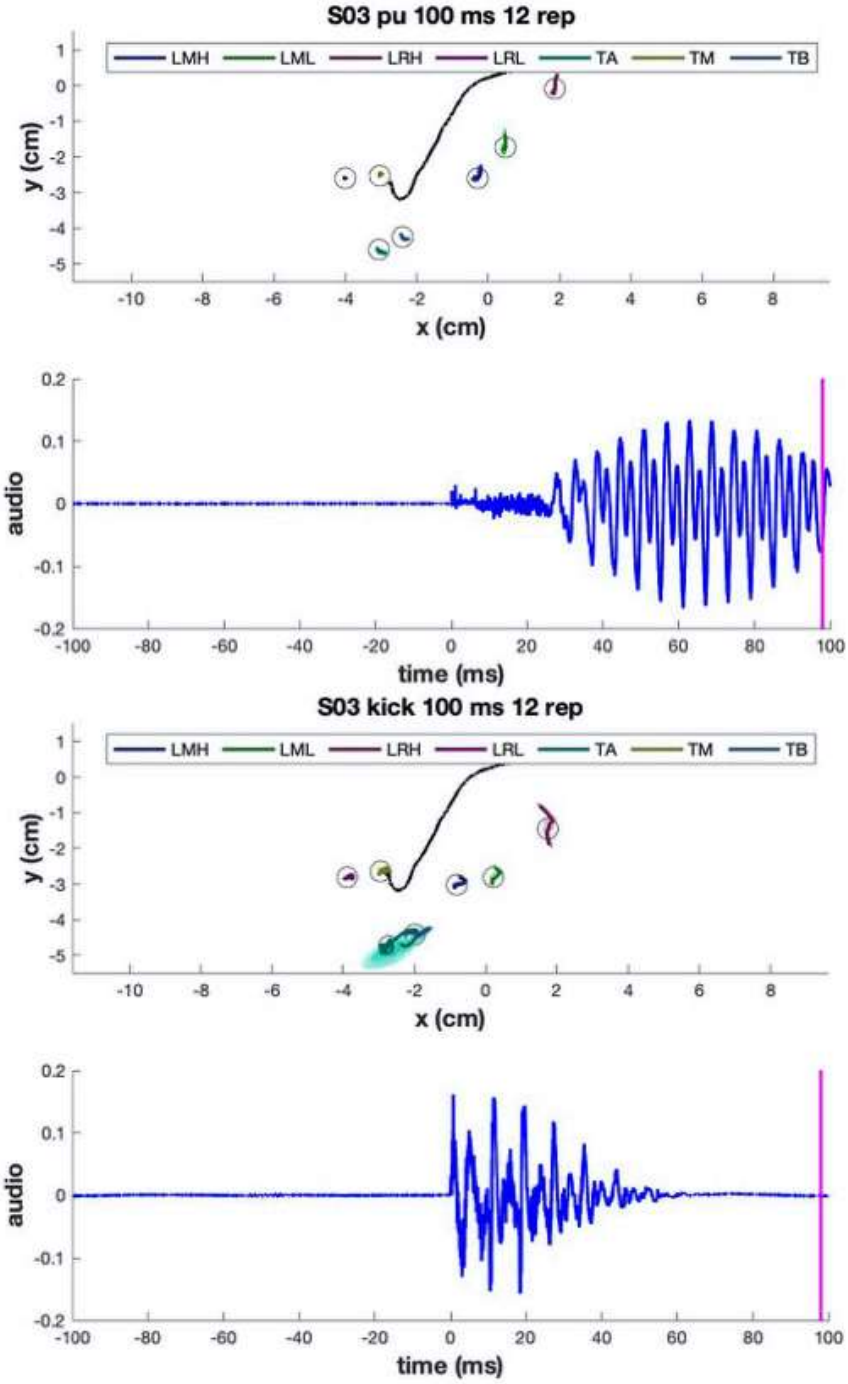


Figure 5.24: Mean and variance of the trajectory of the coils on the lips and tongue, and audio signal of a representative token produced by S03. Top: speech, [pu]; bottom: HBB, Kick. Black solid line: palate contour.

These results confirm the observations made previously in the case of a single beatboxer

(sections 5.3.1 and 5.3.2). The beatboxed consonantal sounds are clearly different from their spoken counterparts, both from an acoustical standpoint (temporal and spectral signature, duration and intensity of the sound) and from an articulatory standpoint (spatial trajectory and articulatory speed). If the place of articulation is the same for the three beatboxed sounds compared to the spoken sounds, differences remain regarding the precise point of articulation, which may be slightly shifted.

The articulatory strategies differentiate beatboxed consonantal sounds from spoken consonants. Bilabial consonants [p] are always released centrally, whereas the bilabial occlusion of the kick can be released laterally. The reason for this may be a better control on lip tension upon occlusion release. This in turn influences the acoustic outcome, allowing to tweak the timbral result. One could see how control of lip tension may be of importance: as our data have shown, lips reach extremely high speeds upon occlusion release and can undergo haphazard displacements, most certainly driven by aerodynamic forces. Involving only the hemilateral portion of the lips in the occlusion release may help reduce this effect. The choice of lateralization side seems variable among beatboxers and not directly related to handedness.

Different articulatory loops are described by the tongue coils, probably related to a different activation strategy of the lingual muscles (Perrier et al., 2003; Thiele et al., 2020). The upward movement of the tongue is observed among all beatboxers, especially in the case of **P**. Both [p] and kick are bilabial sounds, i.e. the main place of articulation is situated on the lips. In both cases the tongue is active as an articulator. While this can be explained in the case of [pu] as the coarticulation of the bilabial plosive with the following close back vowel [u], in HBB the plosive kick sound is not followed by any vocalic sound, yet the tongue displays regular and considerable upward and to a lesser degree forward displacements. This may be explained as related to an ejective production, i.e., a glottalic initiation mechanism. The use of this kind of articulatory mechanism has already been attested in most of the available literature on HBB (Blaylock et al., 2017; De Torcy et al., 2014; Dehais Underdown et al., 2019; Patil et al., 2017; Proctor et al., 2013; Sapthavee et al., 2014) and may serve the purpose of increasing sound efficiency. The upward and forward movements of the tongue could be consequence of the upper motion of the larynx necessary for an ejective production to take place or the tongue may be actively pulling the larynx to produce a more effective ejection, as suggested by Proctor et al. (2013), view that our data seem to endorse (see chapter4).

Different articulatory behaviors correspond to acoustic signatures that clearly distinguish beatboxed sounds from their spoken counterparts. The intensity values of **P** measured in our study are higher than that reported by Dehais Underdown et al. (2019), despite their beatboxer being a participant in our study as well (S04). They indicate an intensity of 68±4 dB for the burst and 60±5 dB for the release noise. However, they don't provide details on the experimental setup, namely the distance between the microphone and the

lips, nor how they calculate intensity. Our higher values may be explained by a different experimental configuration or a somewhat different calculation methodology. Indeed, we measured the intensity over the whole duration of the sound, whereas Dehais Underdown et al. (2019) measured separately the intensity of the burst and that of the release noise. In general, beatboxed sounds are significantly more intense than their spoken counterparts. This is consistent with a glottalic initiation mechanism: ejectives are known to be more intense than their pulmonic equivalents (Ladefoged & Maddieson, 1996). The beatboxed sounds are also longer than the spoken consonants. The coarticulation between consonant and vowel in the spoken sequences certainly contributes to this difference. In HBB sequences, the consonantal sounds are coarticulated with each other, without intertwining with vocalic configurations. Acoustically, they can therefore develop more freely. It would be interesting to verify if this difference in duration abides in a less natural, but phonetically more balanced task of repeating the only spoken consonant, without coarticulation with the vowel. This task would also allow a better comparison of articulatory speeds. In general, HBB seems to require the articulators to move faster than speech, especially in the region where the occlusion occurs. However, in correspondence with the place of articulation, articulatory speeds are lower in speech than in HBB in a phonetic context where the tongue does not have to move much to reach the vowel position ([pu], [ti]), but become higher when the tongue has to move farther distances to reach its final vowel configuration ([ka]). Could this difference in articulatory speed be intrinsically linked to the mechanisms of HBB production or does it depend on the phonetic context?

In conclusion, despite a possible common root, which is reflected in a similar place of articulation, the production of the plosive HBB sounds here explored is clearly distinguished from that of the plosive sounds of speech.

5.4.2 Breathing behavior

The inspection of the breathing data revealed different behaviors among beatboxers and items. In general, syllable repetition and single boxeme repetition tasks showed more variability. Data illustrated in Figure 5.25 confirme that all the beatboxers used a typical strategy during speech: a brief inhalation phase evidenced by an increase of the thoracic and abdominal circumferences followed by a long decrease during sound production indicating a rather regular exhalation.

An exception is S03 production of the /ka/ item (Fig. 5.26 top left): between one syllable and the other, the thoracic circumference increases and then decreases during the acoustic production of each syllable. This suggests that S03 inhaled a small volume of air between each syllable.

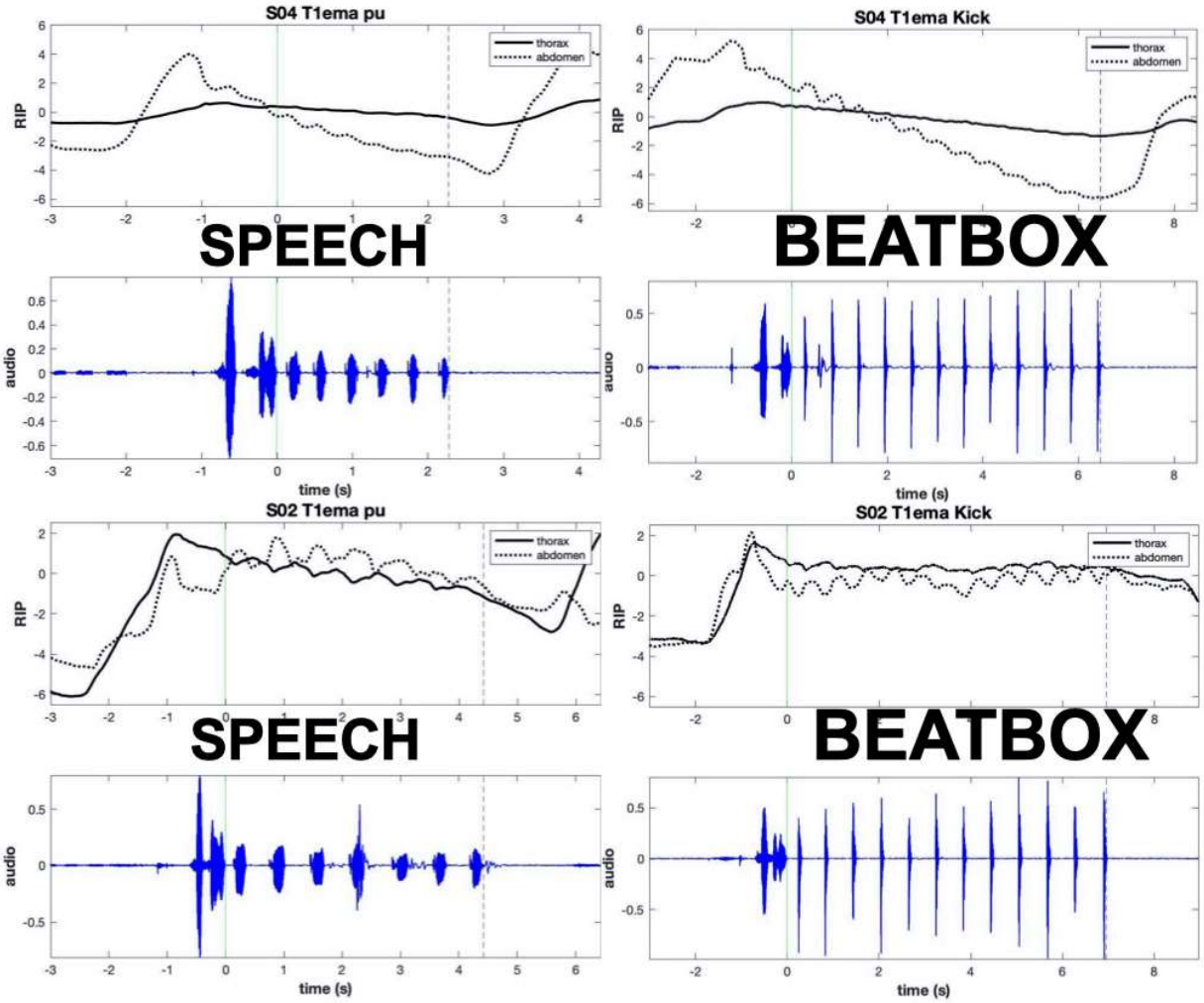


Figure 5.25: Breathing signals of subjects S04 (top) and S02 (bottom). y-axes are arbitrary scales.

Despite a rather common breathing behavior during speech tasks, during HBB tasks (Fig. 5.25) the strategies varied. Some beatboxers such as S04 used a similar behavior to speech, even though more local variations were present in correspondence with the acoustic signal, while the breathing behavior of others such as S02 was peculiar to HBB and characterized by minimal thoracic and abdominal variations during HBB sound production. A noticeable behavior was that of S05 during the rimshot task (Fig. 5.26 bottom right) where the thoracic and abdominal circumferences show opposite attitude. The thoracic circumference increases during the production of the sound, while the abdominal circumference decreases; then, the thoracic circumference decreases and the abdominal circumference increases between sounds. This suggests that S05 produces his K via an ingressive airstream. A positive variation in thoracic circumference is visible along task

production, indicating an overall expansion of the thoracic compartment during phonation. In contrast, no noticeable difference is evident in abdominal circumference, indicating a stabilization of the abdominal compartment likely used as support. A deep exhalation occurs at the end of the task, possibly indicating a situation of "too much air", as already discussed for PS (see Sec.5.3.3). However, these trends seem highly beatboxer and item dependent, as inter-subject and inter-stimuli variability was observed.

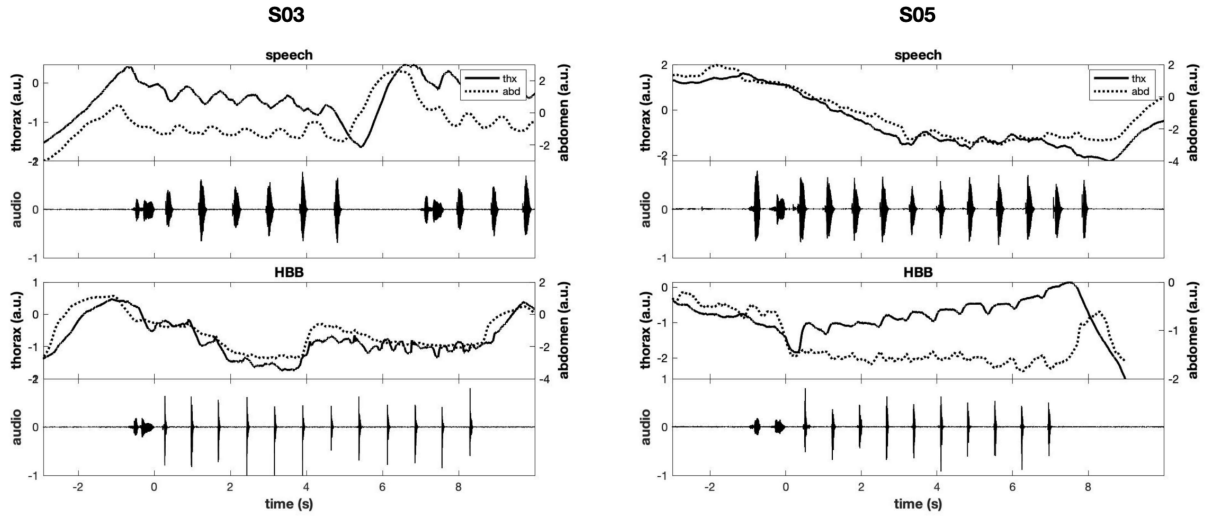


Figure 5.26: Breathing signals of subjects S03 (right) and S05 (left) relative to /ka/ and rimshot (K) items. y-axes are arbitrary scales.

Longer items, i.e., /putikati/ for speech and PtKt for HBB, displayed more common trends (Fig. 5.27). In speech, different contributions of the thoracic and abdominal compartment were observed. An early engagement of the abdomen at the beginning of phonation is visible as a peak in the abdominal curve (S01, S02, S04). An important observation regarding speech tasks (/putikati/) is that there were two ways of producing the task. The most common was to pause every two repetitions or a cluster of /putikati/ (S01, S02, S04, S05). This was protocol induced, in that the pause was requested by the experimenter. However, pauses could be shorter (S01, S05) or longer (S02, S04). In contrast, S03 did not pause and produced all the repetitions in a row. In fact, he produced each syllable disjointed from the adjacent ones, resulting in an unnatural speech flow, very different from connected speech. Of course, the different ways of producing the task have implications on breathing behavior. S02 and S04 breathing behavior shows clear indication of breath group (inspiration followed by phonation on expiration). Despite briefly pausing between clusters, S01 breathing curves do not clearly indicate any noticeable air intake during phonation. This suggests that he produced the whole task on one breath group. S05 only paused briefly between clusters as well, however breath groups are clearly identifiable, albeit of variable duration. S03 shows no uniform behavior, but breath groups are

identifiable: a longer breath group was used in the first and last parts of the task, and three shorter breath groups are visible in the middle part. HBB production was carried out using a clearly different breath strategy than speech, characterized by thoracic and sometimes abdominal hold on a shorter (S02, S04) or longer (S05, but also S01) time scale. S02 and S04 performed an exhalation and an inhalation (drop and increase in circumferences) every two repetitions of PtKt during the pause, even though exhalation is not always present (S02 between the third and fourth clusters, see below for discussion). During sound production of a cluster, S02 shows hold of both thoracic and abdominal circumferences, whereas S04 performs variations in a very limited range around a fixed value. For longer holds, such as S05, two phases are visible. In the first half of the task, an overall hold of thoracic and abdominal circumferences is established, with greater contribution of the thoracic compartment. We may note that in speech tasks S05 shows more circumference variations in the thoracic compartment, and a tendency to stabilize the abdominal circumference, indicating that S05 may have a tendency to use his abdominal compartment as support when phonating. In the second half of the task, the thoracic circumference slowly decreases with local variations, and so does, to a lesser extent and later, the abdominal circumference. Again, S05 uses an ingressive K. The air admission seems more prominent in the thoracic compartment (positive local variations of thoracic circumference). These small but frequent air intakes exploiting the ingressive airstream of the rimshot boxeme seem to justify the absence of audible pauses for silent inspiration, and therefore the need for interruption of the acoustic production for breathing purposes. This behavior is used by S04 as well for the production of his K. Every cluster shows two local, but evident variations of the thoracic and the abdominal circumferences in correspondence with the acoustic signal of K (third and seventh acoustic signal in each cluster, for more in depth discussion, see below).

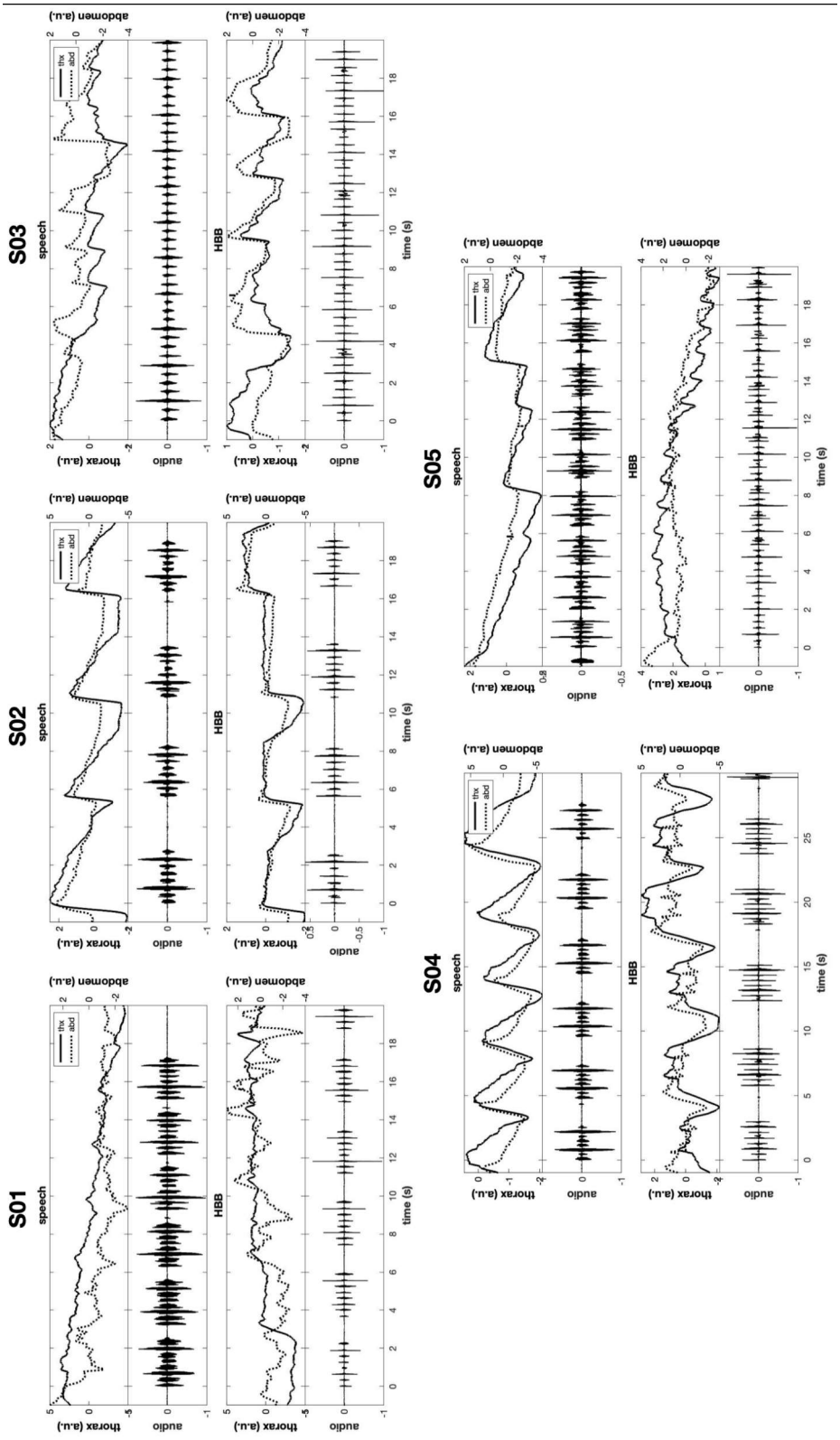


Figure 5.27: Breathing signals of subjects S04 (top) and S02 (bottom). y-axes are arbitrary scales.

In accordance with the observations made on the exploratory data (section 5.3.3), the variations in thoracic and abdominal circumferences were always negative for speech tasks among all beatboxers (Fig. 5.29). Interestingly, S05 is very consistent in his circumference variations (limited interquartile range), even though his air intakes were not evenly distributed along the task. In general, similar variability was observed between speech and HBB tasks, with the noticeable exception of S02 and S05. In fact, S02 shows a very consistent behavior in HBB with very little variation of both the abdominal and more so the thoracic circumferences. This lack of variation suggests that no air is being set in motion by the lungs during the whole cluster, and therefore the initiation mechanisms used must be either glottalic or velaric while the beatboxer is in apnea. This is clarified by the inspection of endoscopic data of equivalent tasks performed by S02 (see sec. 3.2.3). The data show that the laryngeal articulator is in a state of aryepiglottic closure during sound production (Fig. 5.28), while performing vertical displacements. This suggests that the initiator of all the boxemes is indeed glottalic egressive. Further, no opening of the vocal folds is visible during a cluster, confirming that S02 is indeed producing his HBB sounds while in apnea.

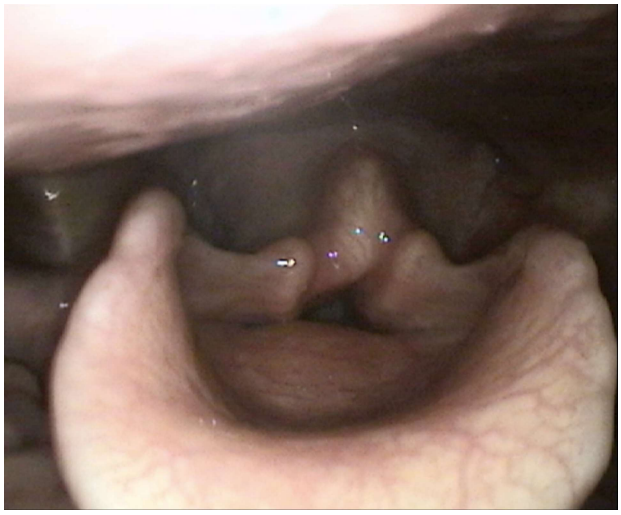


Figure 5.28: Larynx behavior during the production of PtKt by S02.

As mentioned, data regarding S01, S04, and S05 indicate that their K is ingressive (increase in thoracic circumference in proximity of time 0.5). All show opposite contribution of thoracic (circumference increase) and abdominal (circumference decrease) compartments, but with different timing. For S01, the circumference variations seem to happen synchronously between thorax and abdomen, whereas for S04, S05 first the abdominal circumference increases, and only subsequently the thoracic circumference increases, while a more (S04) or less (S05) pronounced decrease happens in the abdominal circumference.

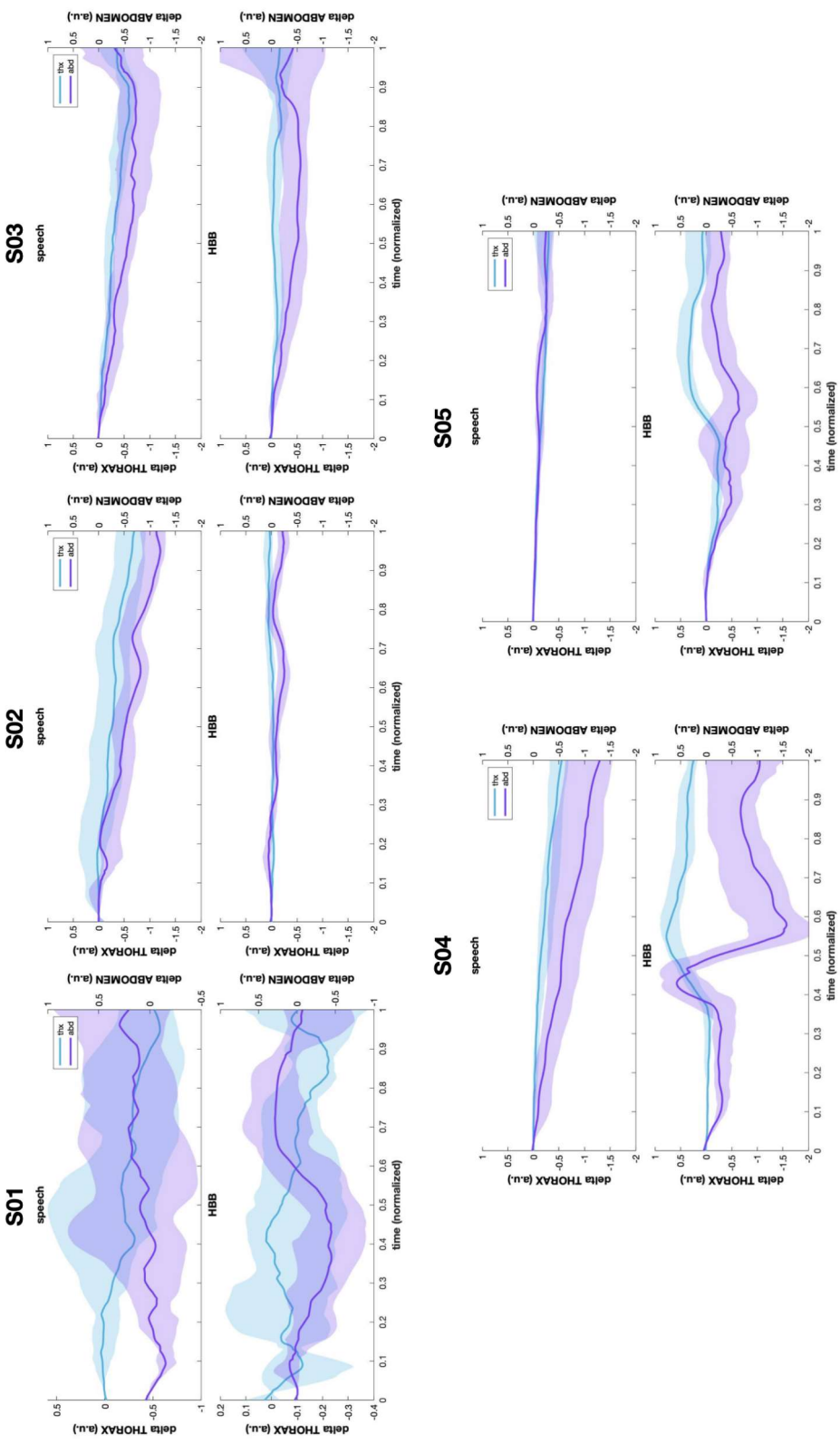


Figure 5.29: Median and interquartile range of breathing signals per repetition. y-axes are arbitrary scales. Time is nomalized.

An earlier increase in abdominal circumference may indicate an expansion of the VT operated by the abdominal compartment to depressurize the air between the lungs and the oral occlusion (see section 5.4.1), and therefore generate the ingressive airstream upon occlusion release.

In summary, intersubject and interstimuli variability notwithstanding, a breathing behavior specific to HBB and different from speech was described. This behavior is characterized by thoracic and sometimes abdominal circumference stabilization, with local variations in correspondence with the acoustic signal of the boxemes, consistent with what observed for PS. The less experienced beatboxer (S02) appeared to beatbox in apnea, whereas the other more experienced beatboxers showed a more skillful breath management, where by means of shallow inhalations and the use of ingressive airstream mechanisms, the sound production can be prolonged over an extended period of time without audible pause specifically dedicated to air intake.

5.5 Beatboxing, it is not speaking

In this chapter, the acoustic, articulatory, and breathing characteristics of 3 HBB sounds were analyzed and compared to those of 3 corresponding consonants of speech in different contexts. The results have confirmed that HBB and speech productions are definitely distinct.

HBB and speech counterparts shared the general region of articulation: [p] and kick are bilabials, [t] and hi-hat are alveolars, [k] and rimshot are velars. However, the articulatory details differed between speech and HBB sounds. In general, the manner of articulation was different. The indication beatboxers give to suppress vowels and emphasize consonants seems to be reflected in our data in a change of initiation mechanism: pulmonic initiation in speech became glottalic initiation in HBB. This, in turn, resulted in more intense HBB sounds, where high acoustic energy is characteristic not only of the burst, but is protracted over the duration of the sound. A glottalic initiation mechanism seemed to also affect the kinematics of the tongue. This is particularly remarkable for the kick, where the tongue is very active, even though the main occlusion is at the lips. More ample movements and higher articulatory speeds, especially in the region of the occlusion, were generally associated with HBB articulations. However it was not possible to definitely determine if this is characteristic of HBB production or is dependent on the phonetic environment. Vowel deletion was confirmed when a CV syllable was converted into a simple HBB boxeme. This resulted in silences between HBB sounds during the occlusion phases and possibly justify the longer duration of HBB sounds with respect to speech consonants. Vowel deletion was also observed when a CVC syllable was converted. This resulted in a boxeme

produced as a double articulation. Breath was used in a typical manner in speech to sustain phonation and breath groups were described. By contrast, our data show that the notion of breath group is not adapted to describe HBB production. Despite intersubject and interstimulus variability, a typical breathing behavior was described, where thoracic and possibly abdominal compartments are generally stabilized and shallow inhalations and exhalations allow for a protracted sound production with no interruption for air intake. Further, ingressive airstreams were observed to serve simultaneously breathing purposes and sound production in two different ways. A same boxeme (e.g., hi-hat) was produced at times via an egressive mechanisms at times via an ingressive mechanism. In this case, the choice of the direction of the airstream seems strictly dependent on breathing needs. Or a boxeme was produced via an ingressive mechanism for aesthetic purposes (e.g., rimshot) and this can be exploited to fulfil breathing needs.

In conclusion, beatboxers may naturally resort to speech sounds to easily set the basis for general indications on place of articulation and source type, by taking advantage of the phonetic knowledge each speaker inherently has. However, substantial modifications at least in initiation mechanism and source location clearly differentiate HBB boxemes and speech consonants. Further, if speech is associated with phonation during the expiratory phase of a breathing cycle and pauses are necessary for air intake, HBB production is characterized by a completely different use of breath, where initiation airstreams simultaneously serve the purpose of sound production and fulfil physiological breathing needs.

Chapter 6

Beatboxing, more than words...

Contents

In this chapter a peculiar way of using the VT is presented, that allows the synchronous production of multiple sounds [1].

6.1 Humming beatboxing, the vocal orchestra within

The most basic way of making music by beatboxing is producing a rhythmic line with sounds that reproduce the drum set sounds, i.e. kick, hi-hat, snare/rimshot, cymbal. More experienced beatboxers, however, can be considered as multivocalists, as they exploit a wide variety of vocal techniques such as rapping, singing, overtone singing, scratching, etc. depending on the style of music they want to produce. In particular, the humming technique can be used to give the impression of multiple sound sources within the same beatboxer: a rhythmic line and a melodic line can be produced simultaneously. This technique is well known by beatboxers and is generally explained as the technique that allows a beatboxer to produce multiple sounds at the same time using the air present in the mouth to produce the rhythm, and the voice to produce the melody. However, how this is achieved remains mostly unexplored from a scientific standpoint. This study focuses on three categories of drum sounds (kick, hi-hat, rimshot) produced as regular HBB

[1] This chapter is based on work presented at two conferences (MAVEBA 2021, CFA 2022). Source: Paroni, A., Loevenbruck, L., Baraduc, P., Savariaux, C., Calabrese, P., and Henrich Bernardoni, N. (2021) Humming beatboxing : the vocal orchestra within. MAVEBA 2021 - 12th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications, Universita Degli Studi Firenze, Dec 2021, Florence, Italy. ⟨hal-03510719⟩

sounds or in the humming technique. Some studies have shown that regular kick, hi-hat, and rimshot are generally produced via a piston-like action of the closed glottis (Blaylock et al., 2017; De Torcy et al., 2014; Proctor et al., 2013), i.e. using a glottalic initiation mechanism (Helgason, 2014). The only published study so far that directly investigates humming boxemes (i.e. HBB sounds) has shown that the humming versions of these three boxemes are produced via a pushing or pulling action of the tongue (Paroni et al., 2021; see Ch. 4), i.e. using a velaric, or more specifically lingual initiation mechanism (Helgason, 2014). The present chapter aims at elucidating the similarities and differences in terms of breathing strategy and articulatory mechanism between regular and humming kick, hi-hat, and rimshot as well as giving some insights on how the vocal tract is configured when producing different sounds simultaneously.

6.2 Material and methods

The results presented in this chapter are drawn from C2.I from 4 beatboxers (S02-S05). Table 6.1 summarizes the participants, the items, and the techniques relative to the data used. For more details, see sec. 3.2.2.2.

Table 6.1: Visual summary of participants, items, and techniques employed.

Participants	**Items**	**Techniques**
S02-S05	/pu/ – kick (**P**) /ti/ – hi-hat (**t**) /ka/ – rimshot (**K**) /putikati/ – **PtKt**	EMA, RIP, EGG, audio, video

The three boxemes kick (**P**), hi-hat (**t**), rimshot (**K**) were produced 12 times each in a row. Each repetition was preceded by [sasɛlə] (English translation: "this is the"). Each boxeme sequence was produced in regular HBB, then using the humming technique. The sequence or *beat* **PtKtPtKt** was repeated 10 times as regular HBB, and 10 times as humming HBB. Spatial trajectories of 9 coils placed on five flesh points of the tongue (apex/blade, middle, right, left and dorsum) and four flesh points of the lips (upper and lower, median and lateral) were extracted from the EMA recordings. Mean trajectories and variance were computed using the commercial software package MATLAB. The segmentations and annotations of the audio signals were used to compute duration and acoustic intensity of the sounds.

6.3 Acoustic, articulatory, and breathing behavior

The three professional beatboxers (S03, S04, S05) produced two versions of the humming boxemes: one was only a sequence of drum sounds, i.e. the rhythmic line (**RL**), the other was a superposition of drum sounds (**RL**) and a hummed melodic line (**ML**). The presence of vocal-fold vibration is attested by the EGG signal in Fig. 6.2. However, one beatboxer alternated vocal-fold vibration and glottal stops when producing his **ML**. The amateur beatboxer (S02) gave only one version of humming boxemes as post-voiced boxemes: no vocal-fold vibration occurred synchronously to the drum-sound production, but was present right after. No vocal-fold vibration was detected during regular HBB production.

Breathing strategies varied among beatboxers and stimuli. Shorter tasks such as boxeme repetition held the most variability. However, a typical pattern emerged during humming **RL** tasks: an increase in thoracic and abdominal circumferences before the initiation of the *beat* was followed by an alternation of decrease and increase during the *beat*. Fig. 6.2 shows that this alternation can be similar to breathing behavior at rest, but was not related to the acoustic outcome. When voicing was added during humming **RL**+**ML** executions, the evolution of thoracic and abdominal circumferences was similar to speech: an increase before vocal-fold vibration initiation attested of air intake, followed by a regular decrease during voicing.

Regular HBB production showed the most varied breathing strategies among beatboxers, especially for shorter tasks (boxemes repetition). Longer tasks, more similar to real-life HBB, attested of a typical behavior: a tendency towards stabilization of thoracic (and possibly abdominal) circumference during the *beat* execution, with small local variations in correspondence with each boxeme acoustic production. In the case illustrated in Fig. 6.2, the more prominent local variations occurred in correspondence with **K** and indicated a rapid increase in thoracic circumference suggesting a small inhalation during the boxeme production.

Articulatory behavior was quite consistent among the four beatboxers for the realization of **P** and **t**, whereas **K** showed more variability. Fig. 6.3 illustrates mean trajectories and acoustic outcomes for S04. **P** was always produced as a bilabial occlusive. However, the tongue was particularly active in the realization of the three variants (Regular, humming **RL**, and humming **RL**+**ML**). As for the two humming versions, the articulatory data show that the superposition of the **ML**, in this case the vocal-fold vibration, to the **RL** did not impact the lingual or labial movements. The tongue was raised high against the palate in the back of the oral cavity and was pushed forward right before and during the occlusion release and the acoustic realization. Breathing data showed no relation between breathing and acoustic production of **P**. In the case of the regular **P**, the tongue assumed a lower position in the oral cavity and underwent an upward displacement that began

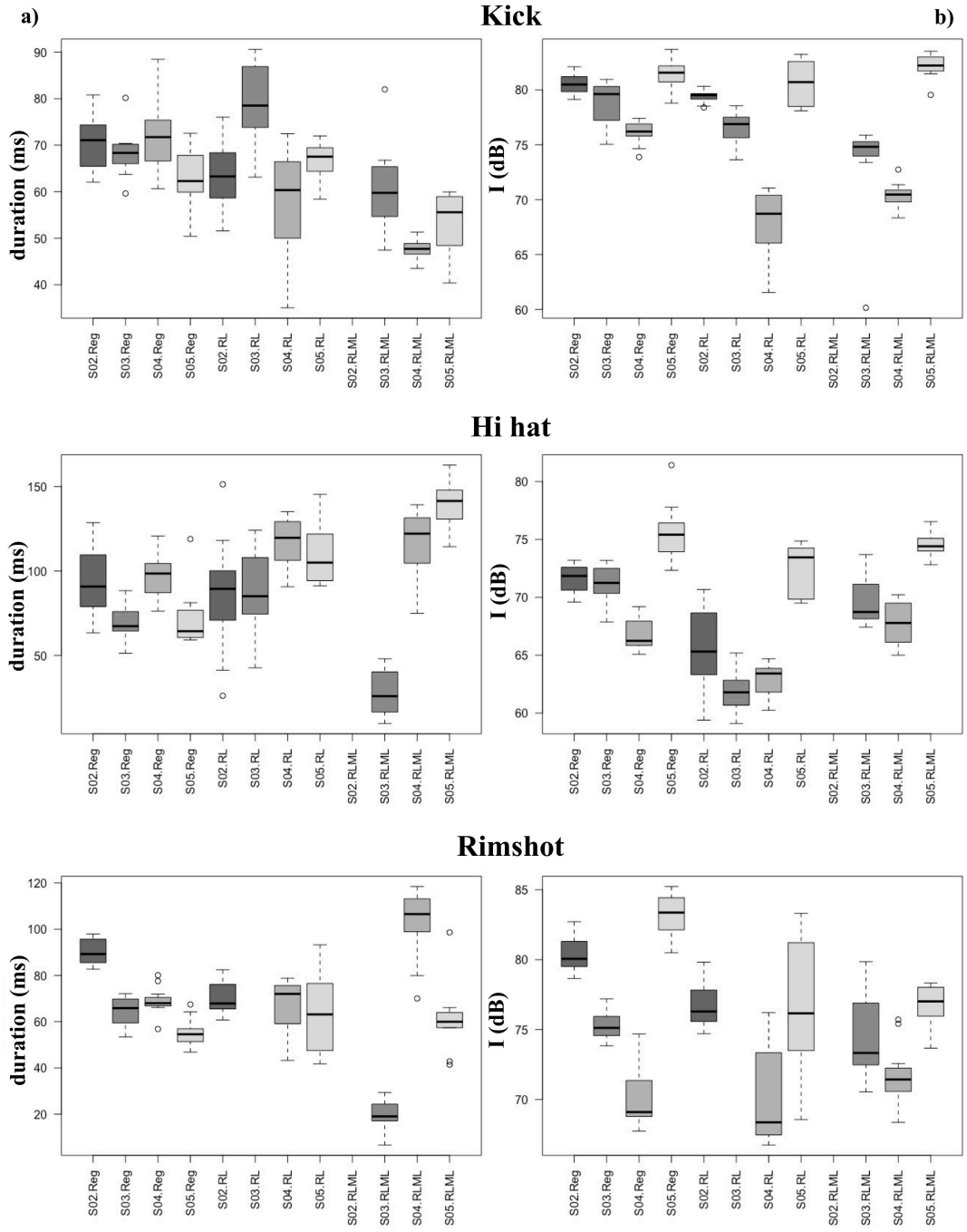


Figure 6.1: Distribution of a) sound duration (in ms) and b) sound intensity (in dB) for each boxeme produced by each beatboxer (S02-S05) as regular HBB (Reg), humming (RL), and voiced humming (RLML).

before the occlusion release and ended after the cessation of the sound. Breathing data showed local decrease in thoracic circumference around sound production, suggesting the use of an egressive airstream. This resulted in slightly shorter and softer humming sounds compared to their regular equivalents (Fig. 6.1). **t** was always produced as an alveolar occlusive. Two main articulatory strategies were observed for the humming versions. One strategy consisted in holding the tongue against the palate and then suddenly producing a rapid downward movement, especially in the middle region, during which the occlusion was released and the sound took place. The other (shown in Fig. 6.3) was via occlusion of the vocal tract in the anterior and posterior region of the oral cavity, creating an air pocket between the middle region of the tongue and the palate and subsequently compressing the trapped air via a pushing action of the middle section of the tongue, releasing the anterior occlusion and producing the sound. In both articulations the tongue assumed a high position, especially in the back region of the oral cavity. The breathing data showed no relation with sound production. Regular **t** was achieved with a lower position of the tongue. Only the more anterior part of the tongue made contact with the palate during the occlusion phase. At occlusion release, the anterior portion of the tongue was lowered and at the same time the posterior portion of the tongue underwent an upward movement. Breathing data showed local decrease in thoracic circumference, suggesting the use of an egressive airstream. The humming versions of **t** generally were longer and softer than the regular **t**. **K** was achieved using the most different articulatory strategies among the beatboxers. For the most part, the humming versions were realized pushing the tongue against the palate and then pulling it down in a rapid motion, while the more anterior region of the tongue was kept in contact with the palate (Fig. 6.3) and the occlusion was released on one side of the tongue. In the humming **PtKt** task, one beatboxer (S05) also produced **K** as a bilabial occlusive, where the pressure buildup was achieved via compression of the cheeks. Again, no relation emerged between breath and acoustic realization. Regular **K** was realized in two different ways. An occlusion was created in the back region of the tongue against the palate, then released via a rapid downward motion of the tongue (Fig. 6.3). In the regular **PtKt** task, two beatboxers used a different articulatory behavior with a different acoustic outcome: the tongue was kept in contact with the palate during the occlusion phase, then the occlusion was released in the back region of the tongue, while the front portion of the tongue was kept in contact with the palate. Breathing data showed a local increase in thoracic circumference, suggesting the use of an ingressive airstream. Once again, humming boxemes resulted as softer and generally slightly shorter than their regular equivalent.

6.4 A peculiar use of the vocal tract

Beatboxers naturally produced two humming versions of **P**, **t**, and **K**: one was the **RL** without **ML**, the other both **RL** and **ML**. These observations suggest that the term "humming HBB" does not imply the presence of a **ML**, but rather the choice of a particular articulatory strategy for the **RL** that is restrained to the oral cavity. This study showed that, while for regular **P**, **t**, and **K** breathing and articulatory behavior are related, with a likely glottalic or pulmonic initiation mechanism in most cases, the humming equivalents systematically switch to a velaric (mostly lingual) egressive or ingressive initiation mechanism. This leads to two main consequences: on the one hand, humming boxemes are generally less intense than regular HBB boxemes; on the other hand, the use of an oral airstream to produce the **RL** allows for the dissociation of breathing and articulation. The high position of the back of the tongue divides the vocal tract into two functional sections that can produce two different sounds at the same time.

In humming HBB, the synchronous production of a rhythmic line and a melodic line is achieved by isolating the oral cavity from the rest of the vocal tract. The oral cavity functions on its own to produce the rhythmic line. Humming kick **P**, hi-hat **t**, and rimshot **K** are produced via velaric (mostly lingual) initiation mechanisms.

This leaves the upstream part of the vocal tract (laryngeal and pharyngeal spaces) available for breathing or producing the melodic line. In the latter case, the humming sound source generated by vocal-fold vibration is propagated into the nasal cavities. This is a skilful and original use of the vocal tract, regularly performed by beatboxers.

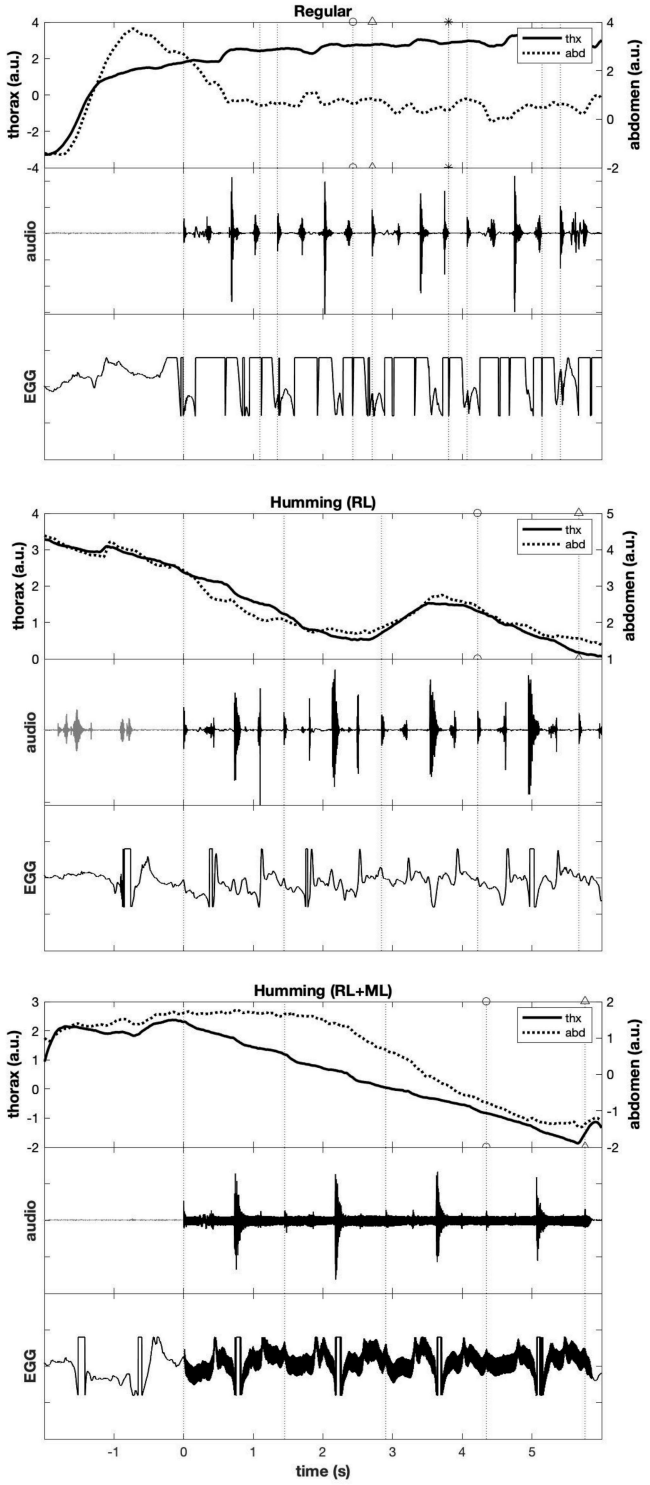


Figure 6.2: Breathing, audio, and EGG signals of S04 producing the beat PtKt. y-axes are arbitrary scales.

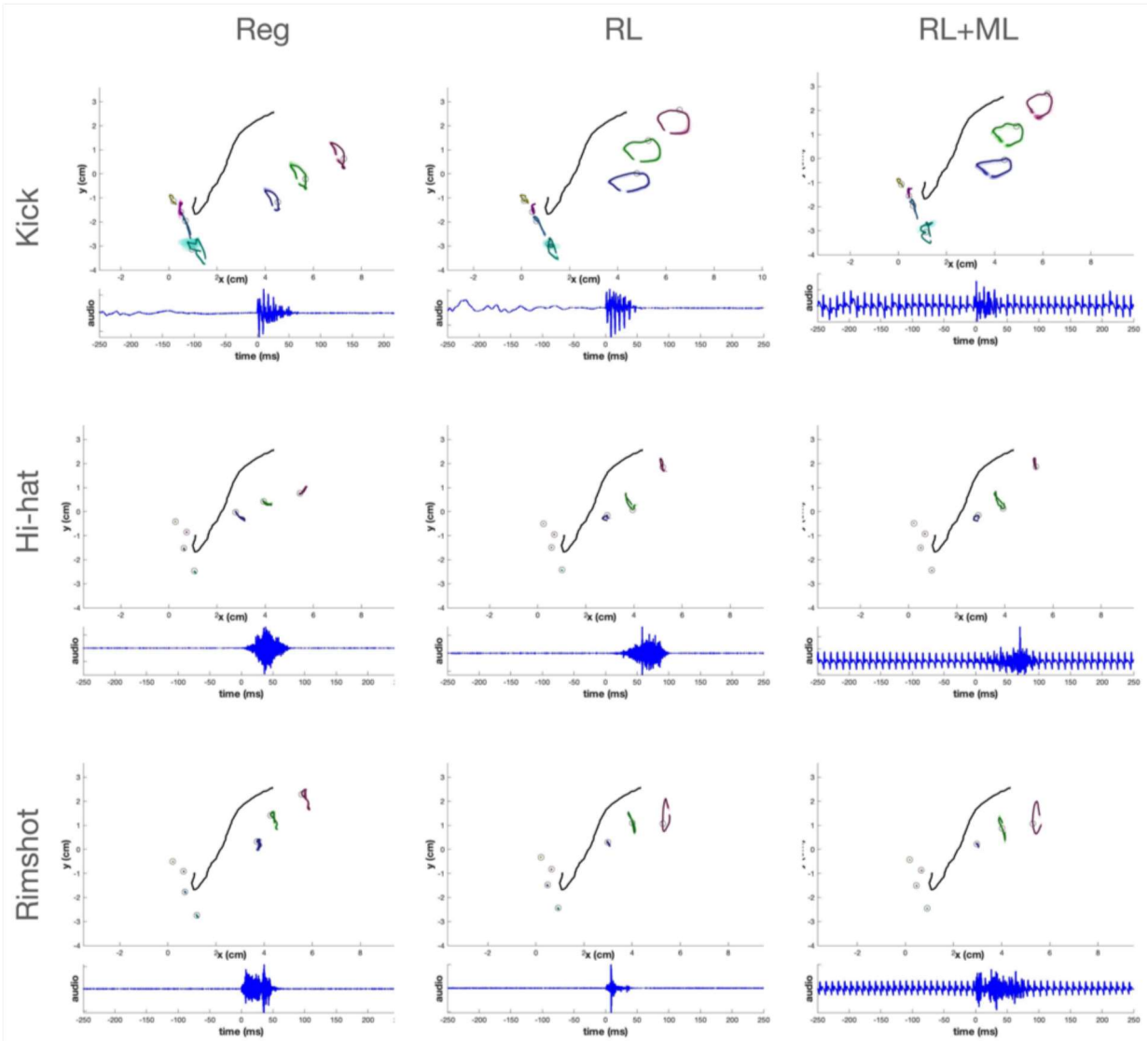


Figure 6.3: Articulatory trajectories of tongue and lip coils in the mid-sagittal plane during regular beatboxing and humming without (RL) or with (RL+ML) melodic line, for singer S04. For each sequence, audio signal of a representative token is plotted. Solid line: palate contour

Part IV

Conclusion

Chapter 7

Conclusion and Perspectives

Throughout this work, we presented our observations on physiological measurements gathered on 6 beatboxers. We exploited synchronized articulatory (EMA), breathing (RIP), electroglottographic (EGG), acoustic, and video data to describe the production mechanisms of 5 categories of drum sounds and 3 speech consonants, focusing on the acoustic characteristics of the sounds, the articulatory and breathing behavior. In particular, we expected to observe:

- a similar acoustic outcome among beatboxers for the same expected sound (e.g., kick);
- a similar articulatory strategy among beatboxers for the same sound (e.g., kick);
- a similar place of articulation for HBB sounds and corresponding speech consonants;
- a different initiation mechanism for HBB sounds and corresponding speech consonants;
- a substantially different breathing behavior in HBB and speech;
- a different initiation mechanism for the same sound produced in regular HBB and in the humming technique;

This chapter is a summary of the main findings of this work.

7.1 Technical remarks

Despite being one of the more common techniques used in speech research, to our knowledge EMA was never used to investigate HBB production. Two main limitations are inherently associated with EMA, that could undermine the feasibility of this technique with regard to HBB production: the presence of wires in the mouth and coils on the articulators, especially the tongue, as well as the adhesion of the glue over time that keeps the coils in place. All

beatboxers reported discomfort and some of them were convinced that they would not be able to beatbox, due to perceived massive perturbation on articulatory movements caused by the presence of the coils and wires. However, the general consensus was that, after a short time for adaptation, beatboxing was in fact possible. Nevertheless, they acknowledged that the acoustic outcome of their sounds was not as good as usual. Concerning the hold of the glue, we experienced relatively few occurrences of detachment. When this was the case, we were able to reattach the coil at the same location, guided by the mark the glue left on the tongue surface. The successful use of EMA is very important to complement the wealth of information other techniques such as MRI and videofibroscopy provide on HBB production mechanisms, inasmuch that it provides quantifiable information on the trajectory and speed of flesh-points. We were able to exploit and integrate these data in a multiple technique approach to disclose articulatory details of drum sounds and better understand the differences with speech productions.

RIP is another widely used technique in speech and singing research that has never been used on HBB. Despite not providing direct information on aerodynamic quantities, we were able to investigate the general breathing behavior of our beatboxers and infer information on airstreams without the inconvenience of a mouthpiece, where we could place EMA coils and sEMG electrodes instead. Further, the use of plethysmographic vests the newer RIP systems provide assures that the sensors are securely kept in place, allowing for a certain freedom of movement for the beatboxer.

In conclusion, EMA and RIP have proven to be viable techniques to gather valuable data for the investigation of HBB production, to complement data provided by other techniques such as endoscopy, MRI, and aerodynamics.

7.2 Results overview and theoretical implications

In Chapter 4 we described the mechanisms underlying the production of 5 categories of drum sounds (kick, hi-hat, rimshot, snare, and cymbals) produced as regular HBB sounds (or 'power', as per PS's terminology) and in the humming technique. We found that each sound was sufficiently different from all the other sounds. In fact, an automatic unsupervised classification was able to distinguish and correctly cluster together the acoustic data, suggesting that each sound had its own acoustic signature. This seems to be relevant, in that it suggests that each sound conveys sufficient acoustic information to be differentiated from the others in a meaningful way. It seems likely that these basic HBB sounds convey a musical meaning as well, be it only the designation of the drum sound they imitate. Thus, we deemed it useful to introduce the notion of *boxeme* to highlight this meaningful aspect of HBB sounds as building blocks of a musical phrase, in analogy with linguistic

phonemes. Of course, more research is needed to investigate the pertinence of this notion and the extent of a possible analogy with linguistic sounds. This goes far beyond the scope of this dissertation, and is left to future work. Despite the limited number of HBB sounds, we were able to describe at least 5 initiation mechanisms: pulmonic egressive and ingressive, glottalic egressive, velaric egressive and ingressive. Two mechanisms (pulmonic ingressive and velaric egressive) are not attested in speech. A similar observation was made by Dehais-Underdown et al. (2021): they were also able to observe all 6 physiologically possible initiation mechanisms on a relatively small set of HBB sounds. Further, we were able to provide articulatory details on the action of the tongue that initiates the velaric mechanisms. From this perspective, we make the choice of designating this initiation mechanism and associated airstream as lingual, as is the choice of Blaylock et al. (2017). Lingual initiation mechanisms were typical of humming variants. Humming sounds were also significantly softer than non-humming sounds. In general, we characterized the drum sound articulations as stop, fricatives, and trills. However, it should be noted that what we designed as a trill using IPA notation is more appropriately described in terms of myoelastic source. The articulatory mechanisms of a bilabial trill designated by the IPA symbol ʘ implies the oscillation of the labial mass under precise aerodynamic conditions and is usually initiated by a pulmonic egressive airstream. In our data, what we observed was a myoelastic vibratory source where a lingual egressive initiation mechanism produced vibration on the mucosa of the lips. To our knowledge, no IPA symbol and/or diacritic can describe this event. The case of the power inward snare is similar. The vibration source visible on the first segment of the sound is likely located at the surface of the lateral rim of the tongue. The presence of tongue movements unrelated to the main place of articulation together with EGG and respiratory considerations pointed to glottalic initiation for all the non-humming sounds with a bilabial occlusion (e.g. kick and snare effects), or alveolar occlusion with egressive airstream (e.g. power closed hi-hat). We observed simple and double articulations on two time scales. In general, shorter sounds were mostly simple articulations, longer sounds were double articulations or sustained sounds, such as fricatives.

In Chapter 5, we investigated the similarities and differences between 3 HBB sounds (kick, hi-hat, rimshot) and 3 French consonants [p, t, k]. We found that the acoustic, articulatory, and breathing strategies substantially differ between HBB and speech. HBB sounds were generally longer and more intense than speech consonants. The longer duration may be related to the absence of vocalic sounds interposed between consonantal sounds. On the other hand, more intense sounds were likely explained by the use of a glottalic initiation mechanism. The general place of articulation resulted to be similar between corresponding speech and HBB sounds, however the precise point of occlusion was not always located in the same place in corresponding HBB and speech sounds. Bilabial occlusions could be lateralized in HBB, the side independent of the handedness of the beatboxer. This strategy is likely to provide better lip tension upon occlusion release and therefore better control

on the timbral properties of the sound, and, if present, the myoelastic source on the lips. More ample and rapid movements, especially at point of occlusion, were observed for HBB articulations. However we could not definitively conclude whether this is HBB-specific or dependent on the phonetic context. In general, we observed similar articulatory strategies among all beatboxers for all 3 speech syllables and for kick and hi-hat, whereas more variety was observed for rimshot[1]. Nevertheless, the place of occlusion was common. This suggests that the reference to speech syllables and/or consonants is useful in that they give general indication on place and manner of articulation, and possibly source type based on inherent phonetic knowledge. However, beatboxers seem to naturally change initiation mechanism and possibly source location, for aesthetic purposes. As De Torcy et al. (2014) observed in the case of a kick, beatboxers think they are producing the same sound as in speech ([p]), just a bit more forcefully, but in fact they add a laryngeal component, i.e., a glottalic initiation, "actually borrowing ejective mechanisms used in other languages of the world". Based on the substantial lingual and labial displacements undisclosed by our data, and in agreement with the considerations made by Blaylock et al. (2017), we would err on the side of caution, as we do not yet know if the details of the ejective mechanism used in HBB, such as for instance magnitude and timing of larynx displacement, are comparable to those used in linguistic ejective consonants. Further, it would be interesting to investigate if there are differences in ejective productions in beatboxers who acquire ejectives in their native language and beatboxers that develop this articulatory mechanism practicing HBB. Moreover, in HBB we observed articulatory loops in absence of coarticulation with vocalic sounds. In the literature (see Weirich et al., 2013 and Thiele et al., 2020) articulatory loops have been attributed to biomechanical properties and muscular effects of the tongue. The fact that we observe articulatory loops, especially during the production of kick, where the tongue should not play an a priori role in the realization of the sound, is an indicator in favour of an active pulling action of the tongue on the larynx that may increase the efficiency of the glottalic initiation mechanism.

Regarding breathing, we observed a typical behavior in speech productions, where inhalation is silent and exhalation is used to sustain phonation. This led to the identification of breath groups. By contrast, the notion of breath group was found to be inapplicable to HBB production. Despite intersubject and interstimulus variability, a typical breathing behavior was described: the thoracic and possibly the abdominal compartments are generally stabilized. A protracted sound production with no interruption for air intake is obtained via the presence of shallow inhalations and exhalations. Further, pulmonic ingressive airstreams were simultaneously exploited for breathing purposes and sound production, and this in two different ways. A same boxeme can be produced at times via an egressive mechanisms, and at times via an ingressive mechanism. In this case, the choice

[1]We did not force the choice of what particular kick, hi-hat or rimshot to produce. The only indication that was given was to produce the classic variant. Inward or outward K are equally common among beatboxers.

of the direction of the airstream seems strictly dependent on breathing needs. Another case of figure is for a boxeme to be produced via an ingressive mechanism for aesthetic purposes, and this can be exploited to fulfil breathing needs. The acquisition of this peculiar breathing behavior may not be trivial and may come with practice. Less experienced beatboxers (S02) may in fact beatbox in apnea.

In conclusion, our work showed that beatboxers may naturally resort to speech sounds to easily provide the basis for general indications on place of articulation and source type, by resorting to the phonetic knowledge each speaker inherently has. However, substantial modifications at least in initiation mechanism and source location clearly take place when speech consonants are rendered into HBB boxemes. Further, if speech is associated with phonation during the expiratory phase of a breathing cycle and pauses are necessary for air intake, HBB production is characterized by a completely different use of breath, where initiation airstreams simultaneously serve the purpose of sound production and fulfil physiological breathing needs.

In Chapter 6, we focus on a peculiar use of the VT that beatboxers exploit to produce a rhythmic line of kick, hi-hat, and rimshot and a melodic line simultaneously. This technique is called 'humming'. We showed that, regular HBB sounds were produced via a glottalic and/or pulmonic initiation mechanism, whereas the humming versions were systematically produced via oral initiation mechanisms (mostly lingual, but also via the pushing action of the cheecks). The exclusive use of oral initiation mechanisms for the rhythmic line and the separation of the oral cavity from the rest of the vocal tract via the contact between the back portion of the tongue and the velar region of the palate dissociates breathing from sound production. Moreover, in the upstream portion of the VT another airstream can be produced, usually pulmonic that can sustain vocal fold vibration, or glottalic. This supplementary airstream can be used for the melodic line. The production of the rhythmic line and of the melodic line are independent of each other, and therefore from a phonetic standpoint the term 'humming' does not indicate the presence of vocal fold vibration propagating through the nasal cavities. Rather, it would imply the choice for the rhythmic line of articulatory strategies that can produce sounds via oral initiation mechanisms, and the possibility to exploit another initiation mechanism for the melodic line. To our knowledge, this is a very skilful use of the VT that is not attested in vocal productions other than HBB. The use humming HBB makes of the VT can be justified in the light of the LAM (Esling et al., 2019), where indeed tongue and larynx are two separate articulators that can operate relatively independently of each other, each in its own segment of the VT, to such an extent that the two can be active at the same time to produce two completely different sounds from two different airstreams. Under this light, the use humming HBB makes of the VT can be regarded as one of the most compelling pieces of evidence in support of the LAM.

The ample use of non speech-like production mechanisms and the particular use of the

VT to generate two airstreams at a time raises the issue of annotating HBB using IPA, which, as already mentioned, is designed for speech sounds. At the end of this work, two perspectives seem reasonable: either extend the IPA symbols to account for the use of, namely, non-linguistic airstream mechanisms and sound sources, the synchronous production of two airstreams, and the different sound sources, or develop a specific annotation system, specifically designed for HBB sounds.

Lastly, if Pike (1943) points out that more sounds can be produced on lung air than with any other mechanism, and in speech the majority of the phones are indeed produced on pulmonic egressive airstream, this does not seem the case of HBB, where pulmonic egressive is only one among the possible initiation mechanisms and probably not even the most common. In fact, if speech heavily relies on phonation and in turn phonation relies on pulmonic egressive airstream, in HBB phonation is only one among the possible ways of producing sound. Pulmonic mechanisms, especially egressive, are well adapted for speech, because the lungs can set in motion large volumes of air, i.e., a pulmonic egressive mechanism can be protracted over time, whereas glottalic and oral, especially lingual, initiation mechanisms can set in motion only a small volume of air, and therefore the airstream thus generated can last only a short time and can serve to produce one or two sounds. The frequent absence of vocalic sounds between boxemes allows for frequent changes in initiation mechanism and/or the frequent repetition of the same mechanism (e.g., during a sequence of drum sounds, one glottalic egressive airstream produces one drum sound). This has another advantage, that of managing breathing in such a way that no pause is needed for air intake. This begs the question of how beatboxers coordinate different initiation mechanisms and how they acquire such a fine control.

7.3 Limitations

It is always important to read the results under the light of the limitations each study inevitably has. This work is not exempt of limitations, of course.

First of all, the number of participants. When we think of studies investigating speech production, it is common to find corpora constituted recording a decent number of speakers. Usually, it is quite easy to find speakers to record. It is far more difficult to find beatboxers willing to come to the lab, often from afar, often for free, and accept to shave their beard "for Science". This explains why we were only able to record 6 beatboxers. Further, on the day we recorded S01, the EMA system was out of order, therefore we were not able to collect articulatory data on him. All the beatboxers were males and all had French as their first language.

In the protocol of C1 we selected the sentence "des petits cookies des gros cookies"

together with "Boots" and "Pâtes", because of their relevance for PS, i.e., PS told us that he used those sentences among others to learn HBB. However, we did not anticipate that PS would delete ə in [pətɨ]. This in turn resulted in an acoustically unexploded [p]. Given that the segmentation was performed on the acoustic data and that all the subsequent analyses were based on the timestamps of the bursts, it was impossible to correctly distinguish between the occlusion release of [p] and that of the following [t]. Repetitions of single HBB sounds and speech syllable are, of course, not ecological conditions. The short duration of the task had a visible impact on breathing behavior, more so in HBB tasks. Longer tasks with repetitions would have the benefit of mitigating this, as well as giving more statistical power to our analyses. The limited number of repetitions we recorded, especially in C2, do not give our statistical analyses, especially those performed on articulatory data, great power, but are informative nonetheless.

Unfortunately, we could not calibrate our plethysmographic system for C2. This means that breathing curves are expressed in arbitrary units. While this provides useful information for the observation of breathing behavior, we could not extrapolate air volume, nor compare the amplitude of thoracic and abdominal circumference variations. Further, as mentioned in section 3.1, RIP measures cumulative variables, among which cross sectional variations that have a muscular nature (e.g., abdominal muscles contraction). While this is classically overlooked in speech and singing research, where the main airstream is pulmonic egressive, it may have more bearing on the estimation of air volumes in HBB, where glottalic initiation is so common and may need a vigorous supporting action from the abdominal compartment.

Initiation mechanisms, especially glottalic, were deduced from indirect observations gathered from the behavior of the oral and glottal articulators and from the breathing behavior. Only in rare cases glottalic initiation mechanisms could be directly observed on endoscopic data. Systematic endoscopic and aerodynamic data would certainly be useful to substantiate our conclusions.

A further limitation of our articulatory analyses concerns transverse (coronal) movements. Because we generally restricted our analyses to the coils placed in the sagittal plane, we only systematically investigated movement and contact of the midline portion of the tongue. Lateral articulation that occurs during HBB may therefore have been missed.

7.4 Future perspectives and recommendations

C2 is a large corpus. Only a small subset of items has been investigated so far, and not even fully. For the most part, the articulatory analyses concerning the tongue were restrained to the coils of the midline. The next step is to complete our observations by investigating

the behavior of the lateral parts of the tongue, particularly in the case of lingual initiation mechanisms.

sEMG data are also under investigation. The preliminary analyses are revealing interesting details on the facial muscle activation in the production of HBB sounds, that seem to further substantiate some articulatory observations, such as longer and more secure occlusion phases in HBB than speech consonants.

The section of the corpus dedicated to the exploration of the production mechanisms of 5 categories of drum sounds (C2.II, see sec. 3.2.2.2) is still to be investigated. This is one of the largest collection of data on HBB drum sounds recorded on multiple beatboxers. It is a unique opportunity to explore and compare production mechanisms among different beatboxers for a given set of sounds. This will further clarify our understanding on whether different beatboxers exploit similar or different production mechanisms to achieve the desired acoustic outcome of a given sound.

As mentioned at the very beginning of this dissertation, this project on the characterization of the production mechanisms of HBB sounds stemmed from the idea that HBB could be a playful, yet effective tool for Speech Therapy, especially in the field of Orofacial Myofunctional Disorders (OMDs). However, we felt the need for caution and a better understanding on the details of this unique vocal art before designing a therapy protocol. This task has proven more complex and time consuming than what we expected. Nevertheless, our results and the recent work of our colleagues around the world certainly substantiate the intuition on the potential of HBB for Speech Therapy. This is certainly a line of research for the future. Yet, it may be useful to remind some brief indications that can be drawn from the results of the present work. First of all, the investigation on breathing behavior has revealed a complex coordination of initiation mechanisms that seems far from trivial to master. The risk for neophytes might be beatboxing in apnea. This in turn could cause muscular tension in the neck region. Further, preliminary results on sEMG data that we could not include in this dissertation have revealed that indeed labial and facial muscular activation is higher in HBB than speech, especially during the occlusion phase of drum sounds. This substantiates our intuition that HBB production is worth investigating more before including it in speech therapy, especially in orofacial myofunctional therapy. Inaccurately designed protocols may induce over-activation of face muscles, particularly the orbicularis oris, in patients where this muscle is already hypertonic.

From a phonetic standpoint, one question that no study has addressed so far is “Do we beatbox the same everywhere in the world?”. In other words, do the production mechanisms of a beatboxer’s first language influence the production mechanisms of HBB? To what extent? And, if this is the case, is there a dependency on the expertise level of the beatboxer or does this persist in experienced and professional beatboxers? Further, as previously discussed, we do not know how the ejective mechanism so widely used in HBB

compares to the glottalic initiation mechanism of the ejective consonants of the world's languages. And lastly, while there seems to be some indication that there are no evident differences in production mechanisms depending on sex (Dehais-Underdown et al., 2021; Patil et al., 2017), there seems to be anecdotal evidence that female beatboxers produce less intense sounds than male beatboxers. Whether this is true at all, if it is due to a general higher level of expertise in male beatboxers or is due to physiological factors remains unexplored from a scientific standpoint.

In conclusion, HBB is really a favourable field for the exploration and comprehension of the production mechanisms of the human voice. We sure hope that more and more scholars will join the effort of understanding this extraordinary vocal art.

Bibliography

Arleo, A. On the phonology of nonsense syllables (U. de Nantes, Ed.). In: In *IIe Journées d'Etudes Linguistiques, " Syllabes "* (U. de Nantes, Ed.). Ed. by de Nantes, U. Nantes, France, 1999, 52–59. https://halshs.archives-ouvertes.fr/halshs-00650629 (cit. on p. 90)

Atherton, M. Rhythm-speak: Mnemonic, language play or song. In: *Proceedings of the international conference on music communication science. sydney, australia*. Citeseer. 2007, 15–18 (cit. on p. 90).

Austin, S. F. (1997). Movement of the velum during speech and singing in classically trained singers. *Journal of Voice*, *11*(2), 212–221 (cit. on pp. 33, 34).

Ball, M. J., Howard, S. J., & Miller, K. (2018). Revisions to the extipa chart. *Journal of the International Phonetic Association*, *48*(2), 155–164 (cit. on p. 82).

Barbier, G., Baum, S. R., Ménard, L., & Shiller, D. M. (2020). Sensorimotor adaptation across the speech production workspace in response to a palatal perturbation. *The Journal of the Acoustical Society of America*, *147*(2), 1163–1178 (cit. on pp. 46, 66).

Bastir, M., García-Martínez, D., Torres-Tamayo, N., Sanchis-Gimeno, J. A., O'Higgins, P., Utrilla, C., Torres Sánchez, I., & Garcia Rio, F. (2017). I n v ivo 3d a nalysis of t horacic k inematics: C hanges in s ize and s hape d uring b reathing and t heir i mplications for r espiratory f unction in r ecent h umans and f ossil h ominins. *The Anatomical Record*, *300*(2), 255–264 (cit. on p. 12).

Benchetrit, G. (2000). Breathing pattern in humans: Diversity and individuality. *Respiration physiology*, *122*(2-3), 123–129 (cit. on p. 91).

Benchetrit, G., Shea, S., Dinh, T. P., Bodocco, S, Baconnier, P, & Guz, A. (1989). Individuality of breathing patterns in adults assessed over time. *Respiration physiology*, *75*(2), 199–209 (cit. on pp. 14, 91).

Besleaga, T., Blum, M., Briot, R., Vovc, V., Moldovanu, I., & Calabrese, P. (2016). Individuality of breathing during volitional moderate hyperventilation. *European journal of applied physiology*, *116*(1), 217–225 (cit. on p. 14).

Blaylock, R., Patil, N., Greer, T., & Narayanan, S. S. Sounds of the human vocal tract. In: *Proceedings of interspeech*. 2017, 2287–2291 (cit. on pp. 39, 41, 45, 66, 83–86, 90, 91, 104, 123, 134, 145, 146).

Boersma, P. (2006). Praat: Doing phonetics by computer. *http://www. praat. org/* (cit. on pp. 59, 67, 92, 93).

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155–180 (cit. on p. 22).

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, *3*, 219–252. https://doi.org/10.1017/S0952675700000658 (cit. on p. 22)

Brunner, J., Hoole, P., Guenther, F., & Perkell, J. S. Dependency of compensatory strategies on the shape of the vocal tract during speech perturbed with an artificial palate. In: *Proceedings of meetings on acoustics 159asa. 9.* (1). Acoustical Society of America. 2010, 060003 (cit. on pp. 46, 66).

Calabrese, P., Besleaga, T., Eberhard, A., Vovc, V., & Baconnier, P. Respiratory inductance plethysmography is suitable for voluntary hyperventilation test. In: *2007 29th annual international conference of the ieee engineering in medicine and biology society.* IEEE. 2007, 1055–1057 (cit. on p. 68).

Calliope, L., & Fant, G. (1989). *La parole et son traitement automatique.* Masson Paris. (Cit. on p. 25).

Cerny, F. J., & Burton, H. (2001). *Exercise physiology for health care professionals.* Human Kinetics Champaign, IL. (Cit. on p. 31).

Contesse, A. (n.d.). Vocal grammatics : Écrire le beatbox (K. Hammou, Ed.). *Conçues pour durer. Perspectives esthétiques sur les musiques hip-hop.* (cit. on p. 39).

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 (cit. on pp. 39, 41, 45, 66, 85, 90, 91, 104, 123, 134, 146).

Dehais Underdown, A., Buchman, L., & Demolin, D. Acoustico-physiological coordination in the Human Beatbox: A pilot study on the beatboxed Classic Kick Drum. In: *Proceedings of the 19th International Congress of Phonetic Sciences.* Melbourne, Australia, 2019, August. https://hal.archives-ouvertes.fr/hal-02284132 (cit. on pp. 66, 85, 90, 104, 123, 124)

Dehais-Underdown, A., Vignes, P., Crevier-Buchman, L., & Demolin, D. In and out: Production mechanisms in human beatboxing. In: *Proceedings of meetings on acoustics 181asa. 45.* (1). Acoustical Society of America. 2021, 060005 (cit. on pp. 41, 45, 91, 104, 145, 151).

Dejours, P, Bechtel-Labrousse, Y, Monzein, P, & Raynaud, J. (1961). Study of the diversity of ventilatory rates in man. *Journal de physiologie*, *53*, 320–321 (cit. on pp. 13, 14).

Dickson, D. R., & Maue-Dickson, W. (1982). *Anatomical and physiological bases of speech.* Little Brown & Company. (Cit. on p. 13).

Dodderi, T., Johnson, A., & Aji, A. M. (2020). Vocal fatigue in beat boxers. *Journal of Voice* (cit. on p. 40).

Draper, M., Ladefoged, P., & Whitteridge, D. (1959). Respiratory muscles in speech. *Journal of Speech and Hearing Research*, *2*(1), 16–27 (cit. on p. 30).

Dromey, C., Heaton, E., & Hopkin, J. A. (2011). The acoustic effects of vowel equalization training in singers. *Journal of Voice*, *25*(6), 678–682 (cit. on p. 34).

Eberhard, A., Calabrese, P., Baconnier, P., & Benchetrit, G. Comparison between the respiratory inductance plethysmography signal derivative and the airflow signal. In:

Frontiers in modeling and control of breathing. Springer, 2001, pp. 489–494 (cit. on p. 68).

Echternach, M., Burk, F., Burdumy, M., Traser, L., & Richter, B. (2016). Morphometric differences of vocal tract articulators in different loudness conditions in singing. *PLoS One*, *11*(4), e0153792 (cit. on p. 34).

Eisele, J., Wuyam, B, Savourey, G, Eterradossi, J, Bittel, J., & Benchetrit, G. (1992). Individuality of breathing patterns during hypoxia and exercise. *Journal of Applied Physiology*, *72*(6), 2446–2453 (cit. on p. 14).

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*, 235–324 (cit. on pp. 20, 31).

Esling, J., Moisik, S., Benner, A., & Crevier-Buchman, L. (2019). *Voice quality the laryngeal articulator model*. Cambridge University Press. (Cit. on pp. 22, 147).

Evain, S., Contesse, A., Pinchaud, A., Schwab, D., Lecouteux, B., & Henrich Bernardoni, N. Reconnaissance de parole beatboxée à l'aide d'un système HMM-GMM inspiré de la reconnaissance automatique de la parole [beatboxed speech recognition using a hmm-gmm system based on automatic speech recognition] (C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, & S. Schneider, Eds.). In: In *6th JEP-TAL-RECITAL Conference, Volume 1 : Journées d'Études sur la Parole [Speech Study Days]* (C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, & S. Schneider, Eds.). Ed. by Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., & Schneider, S. Nancy, France: ATALA, 2020, 208–216. https://hal.archives-ouvertes.fr/hal-02798538 (cit. on pp. 41, 86)

Fabre, C. (2018, June). *Les sons percussifs : des consonnes plosives au Human Beatbox, corrélations acoustiques, aérodynamiques et endoscopiques* (Master's thesis). Service d'ORL et Chirurgie Cervico-Faciale, CHU de Grenoble, Avenue Maquis du Grésivaudan, 38700 La Tronche, France. https://dumas.ccsd.cnrs.fr/dumas-01888799. (Cit. on p. 58)

Fant, G. (1970). *Acoustic theory of speech production*. Walter de Gruyter. (Cit. on p. 14).

Fant, G. (1981). The source filter concept in voice production. *STL-QPSR*, *1*(1981), 21–37 (cit. on p. 20).

Friberg, A., Lindeberg, T., Hellwagner, M., Helgason, P., Salomão, G. L., Elowsson, A., Lemaitre, G., & Ternström, S. (2018). Prediction of three articulatory categories in vocal sound imitations using models for auditory receptive fields. *The Journal of the Acoustical Society of America*, *144*(3), 1467–1483 (cit. on pp. 32, 41).

Fukuda, M., Kimura, K., Blaylock, R., & Lee, S. J. (2022). Scope of beatrhyming: Segments or words (cit. on p. 90).

Gregg, J. W., & Scherer, R. C. (2006). Vowel intelligibility in classical singing. *Journal of Voice*, *20*(2), 198–210 (cit. on p. 34).

Helgason, P. Sound initiation and source types in human imitations of sounds. In: *Proceedings of fonetik*. 2014, 83–88 (cit. on pp. 20, 32, 41, 134).

Henrich Bernardoni, N. Physiologie de la voix chantée: vibrations laryngées et adaptations phono-résonantielles (R Garrel, B. A. de la Bretèque, & V. B. Collectif, Eds.) [Format Broché, 155 pages ISBN 978-2-84023-772-3]. In: In *40èmes Entretiens de Médecine physique et de réadaptation* (R Garrel, B. A. de la Bretèque, & V. B. Collectif, Eds.). Ed. by Garrel, R, de la Bretèque, B. A., & Collectif, V. B. Echanges en réadaptation. Format Broché, 155 pages ISBN 978-2-84023-772-3. Montpellier, France: Sauramps Médical, 2012, March, 17–32. https://hal.archives-ouvertes.fr/hal-00680692 (cit. on p. 16)

Herbst, C. T. (2020). Electroglottography–an update. *Journal of Voice*, *34*(4), 503–526 (cit. on p. 47).

Heyne, M., & Derrick, D. Some initial findings regarding first language influence on playing brass instruments. In: *Proceedings of the 15th australasian international conference on speech science and technology*. 2014, 180–183 (cit. on p. 90).

Heyne, M., & Derrick, D. The influence of tongue position on trombone sound: A likely area of language influence. In: *Proceedings of the 18th international congress of phonetic sciences (icphs)*. University of Canterbury. New Zealand Institute of Language, Brain & Behaviour, 2015 (cit. on p. 90).

Heyne, M., Derrick, D., & Al-Tamimi, J. (2019). Native language influence on brass instrument performance: An application of generalized additive mixed models (gamms) to midsagittal ultrasound images of the tongue. *Frontiers in Psychology*, *10*, 2597 (cit. on p. 90).

Hieronymus, J. L. (1993). Ascii phonetic symbols for the world's languages: Worldbet. *Journal of the International Phonetic Association*, *23*, 72 (cit. on p. 67).

Himonides, E., Moors, T., Maraschin, D., & Radio, M. Is there potential for using beatboxing in supporting laryngectomees?: Findings from a public engagement project. In: *Proceedings of the sempre met2018: Researching music, education, technology. 2018*. International Music Education Research Centre (iMerc) Press. 2018, 165–168 (cit. on p. 40).

Hixon, T. J. (1973). Respiratory function in speech. *Normal aspects of speech, hearing, and language*, 73–125 (cit. on p. 30).

Hixon, T. J., Weismer, G., & Hoit, J. D. (2018). *Preclinical speech science: Anatomy, physiology, acoustics, and perception*. Plural Publishing. (Cit. on pp. 18, 19).

Hoit, J. D., Plassman, B. L., Lansing, R. W., & Hixon, T. (1988). Abdominal muscle activity during speech production. *Journal of Applied Physiology*, *65*(6), 2656–2664 (cit. on p. 30).

Hollien, H., Mendes-Schwartz, A. P., & Nielsen, K. (2000). Perceptual confusions of high-pitched sung vowels. *Journal of Voice*, *14*(2), 287–298 (cit. on p. 34).

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(3), 346–363 (cit. on pp. 68, 94).

Huber, J. E., & Stathopoulos, E. T. Speech breathing across the life span and in disease. In: *The handbook of speech production*. West Sussex, United Kingdom: John Wiley & Sons, 2015, pp. 13–33 (cit. on pp. 12, 13, 16, 30, 31, 91).

Icht, M. (2019). Introducing the beatalk technique: Using beatbox sounds and rhythms to improve speech characteristics of adults with intellectual disability. *International journal of language & communication disorders*, *54*(3), 401–416 (cit. on p. 40).

Kapur, A., Benning, M., & Tzanetakis, G. Query-by-beat-boxing: Music retrieval for the dj. In: *Proceedings of the international conference on music information retrieval*. 2004, 170–177 (cit. on pp. 39, 40).

Ladefoged, P., & Disner, S. F. (2012). *Vowels and consonants*. John Wiley & Sons. (Cit. on pp. 20, 22, 24, 26, 27, 29).

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell Oxford. (Cit. on pp. 20, 22, 27–29, 41, 84, 124).

Lamkin, L. L. An examination of correlations between flutists' linguistic practices and their sound production on the flute. In: *Proceedings of the conference on interdisciplinary musicology*. 2005 (cit. on p. 90).

Leanderson, R., & Sundberg, J. (1988). Breathing for singing. *Journal of Voice*, *2*(1), 2–12 (cit. on pp. 14, 47, 91).

Leanderson, R., Sundberg, J., & von Euler, C. Effects of diaphragm activity on phonation during singing. In: *Transactions of the 13th annual symposium on care of the professional voice*. The Voice Foundation New York. 1984, 165–169 (cit. on pp. 30, 47).

Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., & Susini, P. (2016). Vocal imitations of non-vocal sounds. *PloS one*, *11*(12), e0168167 (cit. on p. 32).

Lemaitre, G., & Rocchesso, D. (2014). On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America*, *135*(2), 862–873 (cit. on p. 32).

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605 (cit. on p. 67).

MacLarnon, A. M., & Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *109*(3), 341–363 (cit. on pp. 16, 30, 91).

Maddieson, I., & Disner, S. F. (1984). *Patterns of sounds*. Cambridge university press. (Cit. on p. 90).

Mathworks: Bioinformatics toolbox: User's guide (r2018b). (2018). (Cit. on pp. 67, 92).

Mignot, P. (2018). *Can human beatbox help cure speech troubles? / can human beatbox serve as a speech physiotherapy method?* http://panamebeatboxhustlers.com/peut-on-reeduquer-les-troubles-de-larticulation-grace-a-la-pratique-du-human-beatbox-fr-en/. (Cit. on p. 40)

Netter, F. H. (2010). Atlas of human anatomy (netter basic science). (Cit. on pp. 15, 17).

Ohala, J. J. The origin of sound patterns in vocal tract constraints. In: *The production of speech*. Springer, 1983, pp. 189–216 (cit. on p. 20).

Ojamaa, R, & Ross, J. Sound and timing must be perfect: Production aspects of the human beatboxing. In: *5th conference on interdisciplinary musicology*. 2009 (cit. on p. 38).

Paroni, A., Henrich Bernardoni, N., Savariaux, C., Lœvenbruck, 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 (cit. on p. 134).

Patel, A., & Iversen, J. (2003). Acoustical and perceptual comparison of speech and drum sounds in the North India tabla tradition: An empirical study of sound symbolism. *Proceedings of the 15th International Congress of Phonetic Sciences* (cit. on p. 90).

Patil, N., Greer, T., Blaylock, R., & Narayanan, S. S. Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging. In: *Proceedings of interspeech*. 2017, 2277–2281 (cit. on pp. 41, 45, 46, 66, 83–85, 90, 91, 104, 123, 151).

Perrier, P., Payan, Y., Zandipour, M., & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *The Journal of the Acoustical Society of America*, *114*(3), 1582–1599 (cit. on p. 123).

Pettersen, V. (2005). Muscular patterns and activation levels of auxiliary breathing muscles and thorax movement in classical singing. *Folia phoniatrica et logopaedica*, *57*(5-6), 255–277 (cit. on pp. 35, 91).

Picart, B., Brognaux, S., & Dupont, S. Analysis and automatic recognition of human beatbox sounds: A comparative study. In: *Proceedings of the 2015 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2015, 4255–4259 (cit. on pp. 39, 41).

Pike, K. (1943). Phonetics. university of michigan publication in language and literature, vol. 21. *Ann Arbor: University of Michigan Press* (cit. on pp. 32, 148).

Pillot-Loiseau, C., Crevier-Buchman, L., Paroni, A., & Bernardoni, N. H. Le human beatbox: D'une utilisation extrême de la voix et de la parole à son utilité en orthophonie. In: *Congrès les phonations*. 2021 (cit. on p. 45).

Pillot-Loiseau, C., Garrigues, L., Demolin, D., Fux, T., Amelot, A., & Crevier-Buchman, L. (2020). Le human beatbox entre musique et parole: Quelques indices acoustiques et physiologiques. *Volume! 16*(1), 125–143 (cit. on pp. 41, 45).

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 (cit. on pp. 39, 41, 45, 66, 83–86, 90, 91, 104, 105, 123, 134).

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/. (Cit. on p. 68)

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, *12*(1) (cit. on p. 46).

Rothenberg, M. (1992). A multichannel electroglottograph. *Journal of Voice*, *6*(1), 36–43 (cit. on pp. 50, 54).

Salomoni, S., Van Den Hoorn, W., & Hodges, P. (2016). Breathing and singing: Objective characterization of breathing patterns in classical singers. *PloS one*, *11*(5), e0155084 (cit. on pp. 35, 91).

Sapthavee, A., Yi, P., & Sims, H. S. (2014). Functional endoscopic analysis of beatbox performers. *Journal of Voice*, *28*(3), 328–331 (cit. on pp. 39–41, 45, 66, 85, 90, 105, 123).

Savariaux, C., Badin, P., Samson, A., & Gerber, S. (2017). A comparative study of the precision of carstens and northern digital instruments electromagnetic articulographs. *Journal of Speech, Language, and Hearing Research*, *60*(2), 322–340 (cit. on pp. 46, 66).

Shea, S., & Guz, A. (1992). Personnalite ventilatoire—an overview. *Respiration physiology*, *87*(3), 275–291 (cit. on p. 14).

Simpson, A. (2014). Ejectives in english and german. *Advances in Sociophonetics, John Benjamins*, 189–204 (cit. on p. 27).

Sinyor, E., Rebecca, C. M., Mcennis, D., & Fujinaga, I. Beatbox classification using ace. In: *Proceedings of the international conference on music information retrieval*. Citeseer. 2005 (cit. on pp. 39, 40).

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of phonetics*, *17*(1), 3–45 (cit. on p. 25).

Stowell, D., & Plumbley, M. D. (2010). Delayed decision-making in real-time beatbox percussion classification. *Journal of New Music Research*, *39*(3), 203–213 (cit. on pp. 84, 85).

Sundberg, J. (1987). The science of the singing voice (dekalb, il, northern illinois press) (cit. on p. 34).

Sundberg, J. (1969). Articulatory differences between spoken and sung vowels in singers. *STL-QPSR, KTH*, *1*(1969), 33–46 (cit. on p. 34).

Sundberg, J. (1974). Articulatory interpretation of the "singing formant". *The Journal of the Acoustical Society of America*, *55*(4), 838–844 (cit. on p. 34).

Sundberg, J., & Skoog, J. (1997). Dependence of jaw opening on pitch and vowel in singers. *Journal of Voice*, *11*(3), 301–306 (cit. on p. 33).

Thiele, C., Mooshammer, C., Belz, M., Rasskazova, O., & Birkholz, P. (2020). An experimental study of tongue body loops in v1-v2-v1 sequences. *Journal of Phonetics*, *80*, 100965 (cit. on pp. 123, 146).

Thomasson, M., & Sundberg, J. (2001). Consistency of inhalatory breathing patterns in professional operatic singers. *Journal of Voice*, *15*(3), 373–383 (cit. on p. 35).

Thorpe, C. W., Cala, S. J., Chapman, J., & Davis, P. J. (2001). Patterns of breath support in projection of the singing voice. *Journal of Voice*, *15*(1), 86–104 (cit. on pp. 35, 91).

Tiede, M., Chen, W.-R., & Whalen, D. (2019a). Fundamental frequency correlates with head movement evaluated at two contrasting speech production rates. *The Journal of the Acoustical Society of America*, *146*(4), 3085–3085 (cit. on pp. 46, 66).

Tiede, M., Mooshammer, C., & Goldstein, L. (2019b). Noggin nodding: Head movement correlates with increased effort in accelerating speech production tasks. *Frontiers in Psychology*, *10*, 2459. https://doi.org/10.3389/fpsyg.2019.02459 (cit. on p. 50)

Titze, I. R. (1980). Comments on the myoelastic-aerodynamic theory of phonation. *Journal of Speech, Language, and Hearing Research*, *23*(3), 495–510 (cit. on p. 16).

Titze, I. R., & Martin, D. W. (1998). Principles of voice production. (Cit. on pp. 10, 12, 30, 31, 34, 91).

Traser, L., Özen, A. C., Burk, F., Burdumy, M., Bock, M., Richter, B., & Echternach, M. (2017). Respiratory dynamics in phonation and breathing—a real-time mri study. *Respiratory physiology & neurobiology*, *236*, 69–77 (cit. on p. 91).

TyTe, & SPLINTER. (2002). Standard Beatbox Notation (SBN). https://www.humanbeatbox.com/articles/standard-beatbox-notation-sbn/. (Cit. on p. 39)

Van den Berg, J. (1958). Myoelastic-aerodynamic theory of voice production. *Journal of speech and hearing research*, *1*(3), 227–244 (cit. on p. 16).

Verma, H., Rana, D., Kumari, A., Dogra, N., et al. (2019). Acoustical and perceptual vocal profile of beatboxers. *Journal of Laryngology and Voice*, *9*(2), 47 (cit. on p. 40).

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395–416 (cit. on p. 67).

Walker, D. C. (1975). Word stress in french. *Language*, 887–900 (cit. on p. 94).

Watson, P. J., Hoit, J. D., Lansing, R. W., & Hixon, T. J. (1989). Abdominal muscle activity during classical singing. *Journal of Voice*, *3*(1), 24–31 (cit. on p. 30).

Weirich, M., Lancia, L., & Brunner, J. (2013). Inter-speaker articulatory variability during vowel-consonant-vowel sequences in twins and unrelated speakers. *The Journal of the Acoustical Society of America*, *134*(5), 3766–3780 (cit. on p. 146).

Yeshoda, K., & Raveendran, R. (2021). Exploring the spectral and temporal characteristics of human beatbox sounds: A preliminary study. *Journal of Voice* (cit. on pp. 41, 45).

Résumé — Le human beatboxing (HBB) est un art vocal en pleine évolution : les beatboxeurs utilisent leurs organes vocaux pour imiter et créer des sons afin de faire de la musique. La clé de cet art est l'expérimentation vocale. Cependant, seules quelques études se sont penchées sur le beatbox jusqu'à présent. Les mécanismes de production et l'étendue de l'exploitation du conduit vocal humain que cet art permet d'atteindre restent largement inexplorés. Si d'une part, les sons de HBB sont produits de manière à ce que l'auditeur naïf ne comprenne pas l'origine humaine de cette production musicale, d'autre part, les beatboxers s'appuient fortement sur les sons linguistiques pour apprendre, enseigner et discuter des boxèmes (sons de HBB). Dans le cadre de ce travail de thèse, nous avons mené une enquête à multiples facettes pour caractériser les mécanismes de production de 5 boxèmes de base (kick, hi-hat, snare, rimshot, cymbale), en mettant en évidence ce qui est spécifique au HBB et ce qui est similaire aux homologues de la parole. Nous avons enregistré 6 beatboxers et analysé les données physiologiques de deux corpus, comprenant des données articulatoires, acoustiques, électroglottographiques, respiratoires et vidéo collectées de manière synchrone, et d'un corpus de données endoscopiques. Nous avons comparé les boxèmes de base kick, hi-hat, rimshot aux homologues parlés [pu, ti, ka] et avons constaté que les trois boxèmes étaient produits comme des articulations occlusives et partageaient le lieu d'occlusion avec les consonnes parlées. Cependant, alors que les consonnes étaient produites par un flux d'air pulmonaire, les boxèmes étaient produits par une action de piston glottique. Ce mouvement laryngé a affecté le mouvement de la langue dans la cavité orale. Des vitesses articulatoires plus élevées lors du relâchement de l'occlusion ont été mesurées pour les boxèmes. Le comportement respiratoire différait entre la parole et le HBB. Pour les tâches de parole, un schéma d'inspiration suivi d'une expiration pendant la production du son a été décrit. Pour les tâches de HBB, un comportement spécifique au HBB a été décrit, avec une tendance à stabiliser la circonférence thoracique et éventuellement abdominale pendant la production du son, et des petites variations locales accompagnant la production acoustique. L'inspiration pendant la production sonore peut être réalisée en passant d'un mécanisme égressif à un mécanisme ingressif pour certains boxemes. Cependant, une variabilité inter-sujet et inter-stimuli a été observée. Nous avons étudié 12 boxèmes appartenant à 5 catégories de sons de batterie (kick, hi-hat, rimshot, snare, cymbales) produits par un beatboxer et avons constaté qu'une classification automatique non supervisée était capable de regrouper correctement les données acoustiques, suggérant que chaque boxème a sa propre signature acoustique. Une variété de gestes articulatoires a été décrite, certains différents de ceux attestés en parole, et une annotation phonétique utilisant l'alphabet IPA a été proposée, soulignant la complexité de la production sonore et les limites de l'annotation basée sur la parole pour les sons de HBB. Deux types de HBB ont été comparés : régulier et humming. La technique du humming permet aux beatboxeurs de superposer une ligne mélodique à une ligne rythmique, alors que le

HBB régulier ne permet de produire que l'un ou l'autre à la fois. Cependant, la manière dont cela est réalisé n'est pas très bien décrite d'un point de vue scientifique. Nous avons constaté que les mécanismes respiratoires et articulatoires sont liés dans le HBB régulier, alors que en humming, ils sont dissociés. La cavité orale est isolée du reste du conduit vocal et fonctionne seule pour produire la ligne rythmique via un mécanisme d'initiation oral. Cela laisse la partie en amont du conduit vocal disponible pour la respiration ou la production de la ligne mélodique. Les vocalisations se propagent à travers les cavités nasales.

Mots clés : phonétique expérimentale, human beatboxing, production vocale artistique.

Abstract — Human beatboxing (HBB) is a rapidly evolving vocal art: beatboxers use their phonation organs to imitate and create sounds to make music. The key to this art is vocal experimentation. However, only a few studies have investigated beatbox sounds so far. The production mechanisms and the extent of paralinguistic exploitation of the human vocal tract this art achieves remain widely unexplored. If on the one hand HBB sounds are produced so that the naive listener does not fathom the human origin of this musical production, on the other hand, beatboxers heavily rely on speech sounds to learn, teach and discuss boxemes (beatbox sounds). Through this thesis work we have conducted a multifaceted investigation to characterize the production mechanisms of 5 basic drum-set boxemes (kick, hi-hat, snare, rimshot, cymbal), highlighting what is specific to HBB and what is similar to speech counterparts. We recorded 6 beatboxers and analyzed physiological data from two corpuses, comprising synchronously collected articulatory, acoustic, electroglottographic, breathing, and video data and from one corpus of endoscopic data. We compared the basic boxemes kick, hi-hat, rimshot to the speech counterparts [pu, ti, ka] and found that the three boxemes were produced as occlusive articulations and shared place of occlusion with the speech consonants. However, where speech consonants were produced via a pulmonic airstream, boxemes were produced via a piston-like action of the glottis. This laryngeal movement affected the motion of the tongue in the oral cavity. Higher articulatory speeds at occlusion release were measured for boxemes. Breathing behavior differed between speech and HBB. For speech tasks, a pattern of air intake followed by exhalation during sound production was described. For HBB tasks, especially the longer tasks, a HBB-specific behavior was described, where a tendency emerged to stabilize the thoracic and possibly the abdominal circumference during sound production, and small local variations accompanied the acoustic production. Air intake during sound production could be achieved by switching from an egressive to an ingressive airstream mechanism

for certain boxemes, such as hi-hat. However, inter-subject and inter-stimuli variability was observed. We investigated 12 boxemes belonging to 5 drum categories (kick, hi-hat, rimshot, snare, cymbals) produced by a beatboxer and found that an automatic unsupervised classification was able to distinguish and correctly cluster together the acoustic data, suggesting that each boxeme has its own acoustic signature. A variety of articulatory gestures was described, some different from those attested in speech, and a phonetic annotation using the IPA alphabet was proposed, highlighting the complexity of sound production and the limits of speech-based annotation for HBB sounds. We contrasted two different kinds of HBB production: regular vs humming. The humming technique allows beatboxers to superpose a melodic line to a rhythmic line, whereas regular HBB allows only for one or the other to be produced at a time. However, how this is achieved is not very well described from a scientific standpoint. Based on our articulatory and breathing data, we found that breathing and articulatory mechanisms are related in regular HBB, whereas in the humming technique they are dissociated. The vocal tract is configured so that the oral cavity is isolated from the rest of the vocal tract and functions on its own to produce the rhythmic line via an oral initiation mechanism where the airstream is generated by tongue or cheek action. This leaves the upstream part of the vocal tract (laryngeal and pharyngeal spaces) available for breathing or producing the melodic line. In the latter case, the humming sound source generated by vocal-fold vibration is propagated into the nasal cavities. This is a skillful and original use of the vocal tract, regularly performed by beatboxers.

Keywords: experimental phonetics, human beatboxing, artistic voice production.

GIPSA-Lab, 11 rue des Mathématiques, Grenoble Campus BP46, F-38402
SAINT MARTIN D'HERES CEDEX