INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Beatboxing Kick Drum Kinematics

Reed Blaylock[1], *Shrikanth Narayanan*[1]

[1]University of Southern California, USA

reed.blaylock@gmail.com, shri@sipi.usc.edu

Abstract

Study of the vocal movements of beatboxing can benefit speech science in a number of ways; but while there are established models of speech motor control that make deterministic predictions about vocal kinematics, little is known about beatboxing motor control. A region of interest method was applied to real-time MRI videos of beatboxing Kick Drums to measure the time to peak velocity—a measurement that is used to assess how well models of speech action predict actual movement trajectories. The average time to peak velocity for Kick Drums is about half of the way (58%) through the total movement duration, similar to the times to peak velocity reported for speech actions. However, while the times to peak velocity for Kick Drums tend to be just above 50%, the times to peak velocity reported for speech sounds are usually a bit below 50%. Further study is needed to assess whether this difference reflects a more extreme constriction goal or a qualitatively different movement pattern.
Index Terms: speech production, human beatbox

1. Introduction

Beatboxers are known for pushing the sound-making potential of the vocal tract to its limits; they create music from a wide variety of vocal sounds, many of which manipulate airflow in ways that are unattested in speech. Research on beatboxing reveals more about the physical capabilities and limitations of the vocal tract, and this in turn may provide valuable information for speech science [1, 2]. Learning how the vocal tract is used by a complex, hierarchical non-speech system like beatboxing provides context for understanding how linguistic sound systems develop and evolve. Speech can also be compared to beatboxing directly in order to learn how much of speech motor control is domain-general (cognitively bound to and unique to the speech system) and how much is shared with other domains like beatboxing.

An outstanding question in beatboxing science is how to model the kinematics of beatboxing sounds, and how those kinematics compare to the kinematics of similarly articulated speech sounds. Common beatboxing sounds like Kick Drums, Closed Hi-Hats, and Inward K Snares use qualitatively similar constrictions to speech sounds—closures of the lips, the tongue tip to the alveolar ridge, and the tongue body to the velum, all followed by explosive releases of air pressure [3, 4]. But whereas the literature on speech motor control has a somewhat comprehensive story for the kinematics of speech sounds, little is currently known about how the vocal tract moves when creating beatboxing sounds. Anecdotally, beatboxing sounds are sometimes thought to be more forceful than speech sounds, a trait which may arise from more rapid movements or larger degrees of compression than speech sounds use. To our knowledge, however, no work has been done assessing whether the beatboxing kinematics are different at all from speech kinematics.

A better understanding of beatboxing vocal kinematics may have both applied and theoretical consequences. Beatboxing is a promising tool for speech therapy [5, 6]. Because some of the basic sounds of beatboxing use the same constrictors (i.e., the lips, tongue tip, and tongue body) and constriction degrees (i.e., full closures) as common speech plosives, the motor control for beatboxing sounds may transfer across domains to speech resulting in stronger speech plosives. More details about how beatboxing movements compare to speech movements could offer the basis for more targeted and effective therapeutic interventions.

Knowing beatboxing kinematics provides the foundations for a theory of beatboxing motor control. In the framework of Tasks Dynamics [7], for example, constrictions in the vocal tract are posited to be governed by differential equations taking the form of point attractors—dynamical systems with a single spatial target, such as a labial closure for a spoken [b] or a narrow constriction of the tongue tip at the alveolar ridge for [z]. These equations have directly observable consequences in the vocal tract: different parameters for the equation yield differences in where the vocal articulators move and how quickly they move at any moment in time (their velocity profile).

In fact, even the literature for speech sounds has not settled on which differential equation is appropriate for speech actions. The original task dynamics equation for a critically-damped mass-spring system is given in Equation 1.

$$\ddot{x} + b\dot{x} + kx = 0 \tag{1}$$

where

$$b = 2\sqrt{k} \tag{2}$$

It predicts that the peak velocity of a vocal tract closure controlled by a gesture will be achieved near the beginning of the gesture's activation (Figure 1), but this is inconsistent with articulatory evidence demonstrating that the time of peak velocity is approximately halfway through a gesture's activation. Alternatives have been suggested which ramp the activation of a gesture [8] or introduce a soft spring term into the equation [9] (Equation 3 depicted in Figure 2), both of which delay the time of peak velocity closer to the midpoint.

$$\ddot{x} + b\dot{x} + kx - dx^3 = 0. \tag{3}$$

10.21437/Interspeech.2023-1880

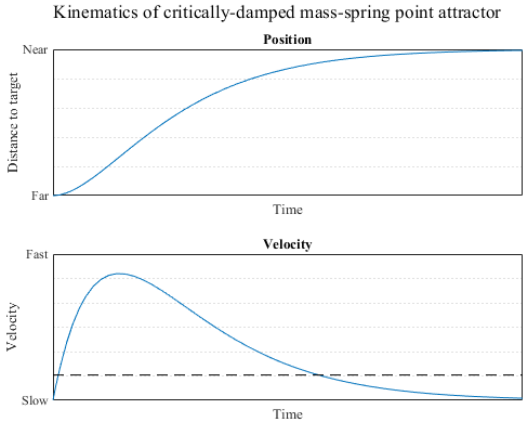

Figure 1: *Position and velocity for a critically-damped mass-spring system (Equation 1). The horizontal dotted line indicates 20% of the peak velocity.*

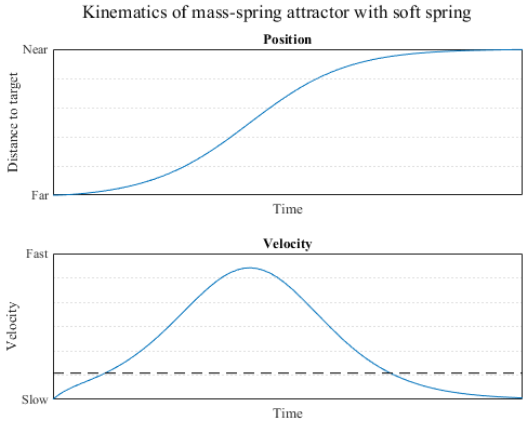

Figure 2: *Position and velocity for a mass-spring system with a soft spring (Equation 3). The horizontal dotted line indicates 20% of the peak velocity*

Beatboxing sounds could likewise be modeled using dynamical point attractors, and this model could be used to predict how more complex beatboxing sounds and patterns will manifest just as extensions and alternatives to Task Dynamics are used to predict how multiple overlapping speech units unfold in time. A prerequisite for such a model of beatboxing is basic kinematic data that can be used to evaluate which differential equation provides the most accurate description of how a beatboxing action is performed.

The broader theoretical question is why speech actions have the control schemes they have. Why these specific point attractors? In Articulatory Phonology [10, 11], the differential equations of Task Dynamics are hypothesized to play the role of phonological gestures: action units that do double-duty as abstract phonological units and concrete articulatory control mechanisms. In this hypothesis, gestures are the result of a dynamical interplay between the communicative tasks of speech and forces of motor efficiency (among other things). But because speech actions are rarely compared against vocal non-speech actions, it is unclear how much the nature of the core dynamical equation is the result of communicative forces or other forces. Is it particularly advantageous for a vocal action to reach its peak velocity halfway through its movement as opposed to later or earlier, or is this simply a property of all vocal point attractors both in speech and out? A fuller accounting of beatboxing kinematics would be the sort of data that could be used to address this question.

The aim of this paper is to begin filling gaps in knowledge about beatboxing kinematics by measuring the time to peak velocity of beatboxing Kick Drums——one of the most fundamental and frequently used beatboxing sounds.

2. Method

2.1. Data acquisition

The data in this study come from a single expert beatboxer, a subset of a larger data set. Two novice beatboxers, one intermediate beatboxer, and two expert beatboxers were asked to produce beatboxing sounds in isolation and in musical rhythms ("beat patterns"), and to speak several passages while lying supine in the bore of a 1.5 T MRI magnet. Skill level designations were given by the intermediate beatboxer who had also contacted the beatboxers, was present for the collection of their data, and provided a beatboxer's insight at several points in the earlier stages of analysis. All participants were cognizant of the nature of the study, provided written informed consent, and were scanned under a protocol approved by the Institutional Review Board of the authors' home institution. Of those five beatboxers, the productions of just one expert are reported in the present study. The two novices and the intermediate beatboxer are not discussed because the aim of this paper is to characterize expert beatboxing, not beatboxing acquisition. (See [12] for an overview of how these beatboxers of different skill levels differ.) Data from the second expert beatboxer are not reported because the beatboxer exhibited large head movements during image acquisition, making kinematic analysis using the region of interest method described below impractical. The beatboxer studied here reported being a monolingual speaker of English.

The beatboxer was asked in advance to provide a list of sounds they know written with orthographic notation they would recognize. During the scanning session, each sound label they had written was presented back to them as a visual stimulus. For each sound, the beatboxer was asked to produce the sound three times slowly and three times quickly, and then to produce the sound in a beat pattern (sometimes referred to hereafter as a "showcase" beat pattern). The beatboxer was also invited to perform beat patterns of their choosing that were not meant to showcase any particular sound. This beatboxer produced over 50 different showcase or freestyle beat patterns.

Data were collected using an rtMRI protocol developed for the dynamic study of vocal tract movements, especially during speech production [13, 14]. The subjects' upper airways were imaged in the midsagittal plane using a gradient echo pulse sequence (TR = 6.004 ms) on a conventional GE Signa 1.5 T scanner (Gmax = 40 mT/m; Smax = 150 mT/m/ms), using an 8-channel upper-airway custom coil. The slice thickness for the scan was 6 mm, located midsagittally over a 200 mm × 200 mm field-of-view; image size in the sagittal plane was 84 × 84 pixels, resulting in a spatial resolution of 2.4 × 2.4 mm. The scan plane was manually aligned with the midsagittal plane of the subject's head. The frames were retrospectively reconstructed to a temporal resolution of 12ms (2 spirals per frame, 83 frames per second) using a temporal finite difference constrained re-

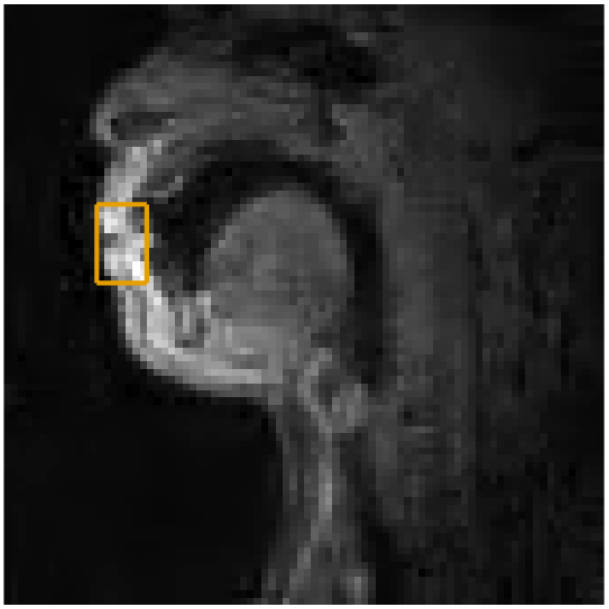

Figure 3: *The closure of a Kick Drum in the region of interest.*

construction algorithm [14] and an open-source library (BART). Audio was recorded at a sampling frequency of 20 kHz inside the MRI scanner while the subjects were imaged, using a custom fiber-optic microphone system. The audio recordings were noise-canceled, then reintegrated with the reconstructed MR-imaged video [15]. The result allows for dynamic visualization and synchronous audio of the performers' vocal tracts.

2.2. Measurements

2.2.1. Pixel intensity time series from region of interest

Time series for lip aperture were created from rtMR video pixel intensities using a region of interest method [16, 17, 18, 19]. A region distills the intensities (brightnesses) of all the pixels it contains into a single mean intensity value. In a video, the region of interest is static but its average pixel intensity changes frame by frame; assembling the frame-by-frame intensity averages into a list creates a time series. With a well-placed region at the lips, changes in average pixel intensity reflect changes in lip aperture: decrease in lip aperture means that the tissue of the upper and lower lips move towards each other and further into the region of interest, thereby increasing the region's overall pixel intensity. The region in this study was manually sized so that the upper and lower lip were just outside the region at their widest aperture (11 pixels tall, 26.4 mm) and so that the region was wide enough to include the full width of the lips during bilabial closures (7 pixels wide, 16.8 mm). Figure 3 shows the closure for a Kick Drum inside the region of interest. The position (but not the size) of the region was adjusted for some videos to account for head movement that occurred between videos during data acquisition.

The pixel intensity time series were low-pass filtered using a second-order butterworth filter with a cutoff frequency of 14 Hz (33% of the Nyquist sampling frequency, 41.5 Hz) run over the data from left-to-right and from right-to-left. Visual inspection concluded that this filter rendered a time series with similar values to the unfiltered intensity and velocity time series while smoothing out higher frequency noise (mean Pearson's r=0.9993 for intensity time series; mean Pearson's r=0.983 for velocity time series).

2.2.2. Acoustic landmarks for each Kick Drum

All beat patterns were manually transcribed into a representation of musical meter by the author based on repeated audiovisual inspection. These transcriptions were used as the basis for labeling events in the acoustic signal captured simultaneously with the rtMR video acquisition. An acoustic event was found for each Kick Drum's release burst using MIR Toolbox (v1.7.2) [20, 21]. Each audio recording was inspected and manually corrected if MIR toolbox found too many or too few acoustic events for a sound, ultimately resulting in a single acoustic time point for each Kick Drum. These events were stored as points on a Praat PointTier [22] and parsed automatically with mPraat [23]. The time of each acoustic event was used as the starting point for searching the pixel intensity time series for changes in lip aperture related to a Kick Drum.

2.2.3. Automatic kinematics extraction

Each sound's acoustic event was used as the basis for automatically finding kinematic moments of interest for a bilabial closure using the DelimitGest function [24] and subsequent automatic corrections. The moments of interest were the time of a local velocity maximum, the time of movement onset, and the time of movement offset. The time of velocity maximum corresponds to the time at which the lips are moving the fastest toward each other; it was found as the moment nearest to the acoustic event of the sound when pixel intensity in the region exhibited the fastest increase. Onset of movement was found where slope of the pixel intensity time series reached 20% of the peak velocity before the local velocity maximum; this corresponds to a time just after the lips have reached a velocity minimum, either because they have switched from opening to closing or because their movement paused for another reason. The same 20% threshold was used to find movement offset after the local velocity maximum, which corresponds to the time shortly before maximum bilabial compression. The time to peak velocity was calculated as the ratio of the duration from time of movement onset to time of velocity maximum and time of movement onset to time of movement offset.

There are several types of Kick Drums in beatboxing, but the ones in this study were limited to the most common variety sometimes known as Classic or "Forced" Kick Drums (B in Standard Beatbox Notation [25, 26]). Articulatorily, this means they were bilabial ejectives [p']. In this paper, the term "Kick Drum" refers only to this specific type of Kick Drum. To minimize the influence of coarticulation, a Kick Drum token was excluded from the analysis if it was co-produced with another sound on the same metrical beat, metrically adjacent to another sound, or observed in post-hoc visual inspection to begin from the release of a bilabial constriction from another sound. In total, 105 Kick Drums produced across 23 showcase beat patterns met these criteria for analysis. Of these, five were not tracked properly by the automatic constriction-finding algorithm and were excluded. Another sixteen were excluded because they were found to have more than one local velocity maximum. This left 84 Kick Drums in the analysis.

3. Result

The pixel intensity time series and associated velocity time series for each labial closure are shown in Figure 4. Red dots

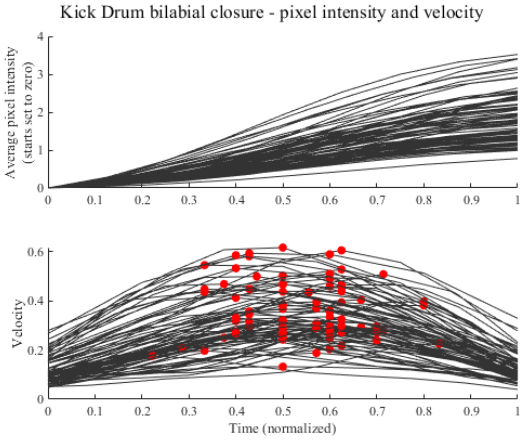

Figure 4: *Pixel intensity (a proxy for aperture) and change in pixel intensity (velocity) for all Kick Drums. Red dots indicate the time of peak velocity for each Kick Drum token.*

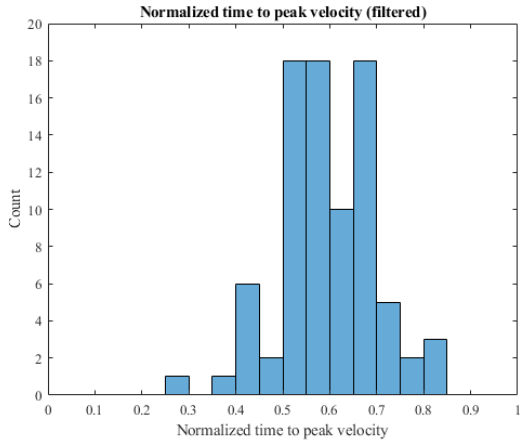

Figure 5: *Histogram of times to peak velocity for all Kick Drum tokens.*

on the velocity time series denote the normalized time of peak velocity for each velocity time series. The overall trend of the curves resembles the velocity trajectory in Figure 2, with most peak velocities between 40% and 70% of the vowel.

Figure 5 shows the distribution of normalized times to peak velocity. On average, peak velocity was achieved 58.8% through the vowel (median=57.1%, mode=66.7%). This indicates that, like the reports for speech data in the literature, the lip closures for Kick Drums are roughly symmetrical insofar as the time of peak velocity is achieved roughly halfway through the time course of the closure.

Figure 6 demonstrates the relationship between movement magnitude and peak velocity: as the magnitude of movement increases, the peak velocity also increases. Sorensen & Gafos [9] argue that while the critically-damped mass-spring system without a soft spring (with or without ramped activation) predicts a fully linear relationship between movement magnitude and peak velocity, the autonomous equation with a soft spring predicts a nonlinear relationship in which peak velocities increase more slowly at higher magnitudes. Assessing whether a distribution has a nonlinear relationship is complicated, but

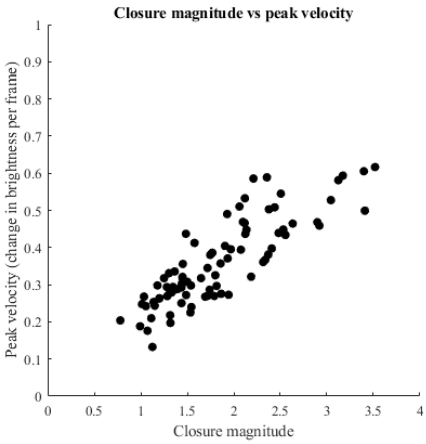

Figure 6: *Scatter plot of constriction magnitude (the overall change in brightness from movement onset to movement offset) against peak velocity (the maximum frame-to-frame change in brightness for each trajectory).*

at a glance the relationship appears fairly linear except perhaps for the peak velocities with magnitude ≥ 3 which may be a bit lower than predicted by a linear relationship. More data with larger closure magnitudes would be needed to examine this further.

4. Discussion

The time to peak velocity measurement of 58% indicates that Kick Drum closures are roughly symmetrical in time: they start slow, increase in speed until about halfway through the sound, then slow back down until the constriction target is achieved. This velocity profile is consistent with the kinematics generated by a differential equation with a soft spring (Equation 3, Figure 2). This is also consistent with the literature for speech movements, indicating that the Kick Drum has a similar movement profile to at least some speech sounds.

The reports in the literature are that times to peak velocity are approximately symmetrical, but usually still earlier than the actual halfway point of the closure. In the most comparable case, Byrd & Saltzman [8] report times to peak velocity for labial closures of a bit under 50%, but other findings for speech closures also report times to peak velocity slightly earlier than halfway. The Kick Drum times to peak velocity, on the other hand, are somewhat later—often 50% or greater.

The fact that the peak velocities are a bit after 50% could indicate the same basic time course as a speech gesture but with a more extreme compression target. In that case the pixel intensities may not register the full magnitude of the compression, which could cut off the movement early—thus making the time of peak velocity later in proportion to the measured movement duration. The Kick Drums could be said to be "forceful" insofar as they have closures caused by a goal for strong compression; however, it is not clear from this study whether that compression is any tighter than the compression for bilabial stops in speech.

5. Acknowledgments

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

6. References

[1] C. Pillot-Loiseau, L. Garrigues, D. Demolin, T. Fux, A. Amelot, and L. Crevier-Buchman, "Le human beatbox entre musique et parole : quelques indices acoustiques et physiologiques," *Volume!*, no. 16 : 2 / 17 : 1, pp. 125–143, Jun. 2020. [Online]. Available: http://journals.openedition.org/volume/8121

[2] A. Paroni, N. Henrich Bernardoni, C. Savariaux, H. Lœvenbruck, P. Calabrese, T. Pellegrini, S. Mouysset, and S. Gerber, "Vocal drum sounds in human beatboxing: An acoustic and articulatory exploration using electromagnetic articulography," *The Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 191–206, Jan. 2021. [Online]. Available: http://asa.scitation.org/doi/10.1121/10.0002921

[3] R. Blaylock, N. Patil, T. Greer, and S. S. Narayanan, "Sounds of the Human Vocal Tract," in *INTERSPEECH*, 2017, pp. 2287–2291. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2017/blaylock17_interspeech.html

[4] A. Dehais-Underdown, P. Vignes, L. Crevier-Buchman, and D. Demolin, "In and out: production mechanisms in Human Beatboxing," in *Proceedings of Meetings on Acoustics*, vol. 45, Seattle, Washington, 2021. [Online]. Available: http://asa.scitation.org/doi/abs/10.1121/2.0001543

[5] M. Icht, "Introducing the Beatalk technique: using beatbox sounds and rhythms to improve speech characteristics of adults with intellectual disability: Using beatbox sounds and rhythms to improve speech," *International Journal of Language & Communication Disorders*, vol. 54, Nov. 2018.

[6] ——, "Improving speech characteristics of young adults with congenital dysarthria: An exploratory study comparing articulation training and the Beatalk method," *Journal of Communication Disorders*, vol. 93, p. 106147, Sep. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0021992421000708

[7] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, Dec. 1989, publisher: Routledge _eprint: https://doi.org/10.1207/s15326969eco0104_2. [Online]. Available: https://doi.org/10.1207/s15326969eco0104_2

[8] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple prosodic boundaries," *Journal of Phonetics*, vol. 26, no. 2, pp. 173–199, Apr. 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447098900717

[9] T. Sorensen and A. Gafos, "The Gesture as an Autonomous Nonlinear Dynamical System," *Ecological Psychology*, vol. 28, no. 4, pp. 188–215, Oct. 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10407413.2016.1230368

[10] C. P. Browman and L. M. Goldstein, "Towards an Articulatory Phonology," *Phonology Yearbook*, vol. 3, pp. 219–252, 1986, publisher: Cambridge University Press. [Online]. Available: http://www.jstor.org/stable/4615400

[11] C. P. Browman and L. Goldstein, "Gestural Structures and Phonological Patterns," Haskins Laboratories, New Haven, Connecticut 06511, Tech. Rep. SR-97-98, 1989.

[12] N. Patil, T. Greer, R. Blaylock, and S. S. Narayanan, "Comparison of Basic Beatboxing Articulations Between Expert and Novice Artists Using Real-Time Magnetic Resonance Imaging," in *Interspeech 2017*. ISCA, Aug. 2017, pp. 2277–2281. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2017/patil17b_interspeech.html

[13] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, Apr. 2004, publisher: Acoustical Society of America. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.1652588

[14] S. G. Lingala, Y. Zhu, Y.-C. Kim, A. Toutios, S. Narayanan, and K. S. Nayak, "A fast and flexible MRI system for the study of dynamic vocal tract shaping," *Magnetic Resonance in Medicine*, vol. 77, no. 1, pp. 112–125, 2017, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.26090. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.26090

[15] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, Oct. 2006, publisher: Acoustical Society of America. [Online]. Available: https://asa.scitation.org/doi/full/10.1121/1.2335423

[16] A. C. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-Driven Analysis of Realtime Vocal Tract MRI using Correlated Image Regions," in *Interspeech 2010*, Makuhari, Chiba, Japan, 2010, pp. 1572–1575.

[17] R. Blaylock, "VocalTract ROI Toolbox," 2021, https://zenodo.org/badge/latestdoi/98065485. [Online]. Available: https://github.com/reedblaylock/VocalTract-ROI-Toolbox

[18] A. C. Lammert, V. Ramanarayanan, M. I. Proctor, and S. S. Narayanan, "Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis," in *Interspeech 2013*, 2013, pp. 959–962.

[19] M. Proctor, A. Lammert, A. Katsamanis, L. Goldstein, C. Hagedorn, and S. Narayanan, "Direct Estimation of Articulatory Kinematics from Real-Time Magnetic Resonance Image Sequences," in *Interspeech 2011*, 2011, pp. 284–281.

[20] O. Lartillot, P. Toiviainen, and T. Eerola, "A Matlab Toolbox for Music Information Retrieval," in *Data Analysis, Machine Learning and Applications*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Berlin, Heidelberg: Springer, 2008, pp. 261–268.

[21] O. Lartillot, P. Toiviainen, P. Saari, and T. Eerola, "MIRtoolbox." [Online]. Available: http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[22] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[23] T. Bořil and R. Skarnitzl, "Tools rPraat and mPraat," in *Text, Speech, and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2016, vol. 9924, pp. 367–374, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-45510-5_42

[24] M. Tiede, "MVIEW: Multi-channel visualization application for displaying dynamic sensor movements," 2010.

[25] D. Stowell, "The Beatbox Alphabet," 2003. [Online]. Available: http://www.mcld.co.uk/beatboxalphabet/

[26] G. TyTe and M. SPLINTER, "Standard Beatbox Notation (SBN)," Sep. 2014. [Online]. Available: https://www.humanbeatbox.com/articles/standard-beatbox-notation-sbn/