Adjusting the Frame: Biphasic Performative Control of Speech Rhythm

Samuel Delalez, Christophe d'Alessandro

Adjusting the Frame: Biphasic Performative Control of Speech Rhythm

Samuel Delalez[1], *Christophe d'Alessandro*[2]

[1]LIMSI, CNRS, Université Paris-Saclay, Bât 508, rue John von Neumann, Campus Universitaire,
F-91405 Orsay, France
[2]Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7190, Institut Jean Le Rond d'Alembert,
4 place Jussieu, F-75005, Paris, France

delalez@limsi.fr, cda@dalembert.upmc.fr

Abstract

Performative time and pitch scaling is a new research paradigm for prosodic analysis by synthesis. In this paper, a system for real-time recorded speech time and pitch scaling by the means of hands or feet gestures is designed and evaluated. Pitch is controlled with the preferred hand, using a stylus on a graphic tablet. Time is controlled using rhythmic frames, or constriction gestures, defined by pairs of control points. The ”Arsis” corresponds to the constriction (weak beat of the syllable) and the ”Thesis” corresponds to the vocalic nucleus (strong beat of the syllable). This biphasic control of rhythmic units is performed by the non-preferred hand using a button. Pitch and time scales are modified according to these gestural controls with the help of a real-time pitch synchronous overlap-add technique (RT-PSOLA). Rhythm and pitch control accuracy are assessed in a prosodic imitation experiment: the task is to reproduce intonation and rhythm of various sentences. The results show that inter-vocalic durations differ on average of only 20 ms. The system appears as a new and effective tool for performative speech and singing synthesis. Consequences and applications in speech prosody research are discussed.

Index Terms: performative synthesis, speech rhythm

1. Introduction

Performative (i.e. real-time controlled) time and pitch scaling is a new research paradigm for prosodic analysis by synthesis. Like in a musical instrument, speech prosody is ”played” or controlled by hands and feet. Such a system is useful for several applications. Expressive speech and singing synthesis was the primary motivation. But the system could also provide new insights to some aspects of brain organization theories about voice production and motor planning. Ultimately, it could provide new speech therapy techniques, with the help of audio-motor synchronization. This raises many fundamental questions on the control gestures, control parameters, and finally prosodic representation. In this analysis-by-synthesis process, a representation is valid if it allows for accurate control or stylization of intonation and rhythm. In a previous study, the effectiveness of ”chironomic” control of intonation stylization has been demonstrated [1]. Intonation contours controlled by drawing gestures, using a stylus on a graphic tablet, were perceptually equivalent to natural contours. In the present paper, the question of rhythmic control is addressed.

The approach defended in this paper is based on the frame/content theory of speech production [2]. It postulates that speech is made of distinctive sounds (phones, or content) contained in syllabic frames. The frames correspond to oscillatory motions of the articulators (particularly jaws), carrying the prosodic rhythm. Syllable frames are opening/closing motions, and can therefore be described by two main points. Following the Greek prosodic terminology, these points are called here ”arsis” (weak beat, merging syllabic attack and coda) and ”thesis” (strong beat, vocalic nucleus).

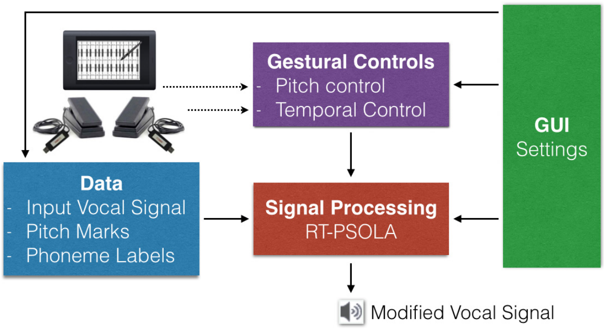

Figure 1: *VOKinesiS - System Overview*

VOKinesiS, a system for real-time rhythm and intonation control based on modification of pre-recorded voice samples has been designed, developped and evaluated. Intonation and vocal effort are controlled using a graphic tablet, like in other performative systems [3, 4, 5]. Articulation timing and rhythm are controlled thanks to syllabic sized chunks manipulation, using various methods, like finger tapping or continuous expression pedal motions. An overview of the VOKinesiS system is displayed in Figure 1. A pre-recorded and labeled (pitch marks and phoneme labels) voice signal is modified with the RT-PSOLA algorithm [6], according to the user's gestures.

The paper is organized as follows: real-time control of the various vocal parameters is presented in section 2, with some emphasis on rhythmic organisation of voice production. Section 3 presents an evaluation of VOKinesiS in a prosodic imitation task. Section 4 is a discussion on syllabic rhythm control and its applications.

2. Controlling prosodic parameters

VOKinesiS computes a synthesis signal by pitch and time scaling of the original signal, according to the pitch and rhythmic targets controlled by the user's gestures. Pitch is controlled directly on a surface. Prosodic rhythms control is achieved with the help of *target time-instant* noted t_i. Pitch and time scaling are achieved using an improved real-time implementation of the PSOLA [7] method. The output signal is synthesised by pitch synchronous re-sequencing of the original pitch periods, according to the target pitch and time-instants. Pitch periods may be duplicated when the signal is lengthened, or may be skipped when it is shortened. Unvoiced parts are processed with a special random duplication method to avoid tonal noises.

2.1. Pitch control

Pitch is controlled by a stylus on a graphic tablet, performing hand gestures similar to writing gestures. The preferred hand is used. This has been used in previous performative synthesis systems [4, 5], and it proved to be a very effective and intuitive pitch control method. The precision and accuracy of hand pitch control is equivalent to (and even better than) intonation control by singers and speakers [1, 8, 9].

2.2. Principles of rhythmic control

Two main rhythmic scales can be considered in speech [10]: intonational rhythm and syllabic rhythm. Intonational rhythm corresponds to intonation contours. It can be managed with a stylus. This has been used in synthesizers using a graphic tablet and only vocalic (or sustained) sounds [4, 5]. When articulation is also considered, intonational rhythm is built on the underlying syllabic rhythm, regardless of the rhythmic organisation (i.e. stress, syllable or mora timed) of the language [11]. For syllabic rhythm description, from the perception viewpoint, the concept of P-center (Perceptual Center) seems useful and relevant. P-centers are able to represent the syllable rhythmic time point. It is located near the vowel onset [10, 12, 13, 14]. A simple way for controlling syllabic rhythm would be to control the instants corresponding to P-centers. But preliminary experiments showed that controlling only P-centers is insufficient for accurate syllable re-sequencing: transients must also be controlled. A new method is designed for accurate and intuitive time-domain manipulation of any prerecorded voice segment, at an intra-syllabic (or phonemic) level of detail. The aim is to be able to control fine articulation timing, but also to ensure naturalness and sound quality. The controlled time scale is of the order of magnitude of syllable components, using two points for a minimum of about $80 - 100ms$. This time scale can be controlled e.g. by finger taping. Using continuous controllers (instead of tapping), it is even possible to decompose this time scale and to control articulation timing (a minimum of about $10ms$).

2.3. Syllabic Control Points

Syllabic rhythm depends on the syllable structure. The syllable is often described with three components : the attack, the vocalic nucleus, and the coda. The attack and the coda correspond to one or more consonants, and the nucleus to the vowel. A syllable always contains a vocalic nucleus, but the attack and the coda are not necessarily present. For the purpose of rhythm production, this definition of the syllable, as a one-to-three phased unit, appeared not well suited.

According to the frame/content theory [2], speech is organized in syllabic frames (cycles of mouth open-close alternation) and segmental content (phonemes), which are controlled separately. The attacks and codas of successive syllables correspond to the opening and closure motions of the vocal apparatus, when the vowels correspond to the open positions.

These cycles of opening and closing can be exploited for rhythmic control. Then the concepts of ”Arsis” and ”Thesis” seemed well suited for performative control. The thesis (derived from the Greek, ”to place, to lower” the foot) represents the stable part of the segment, in our case the vowel or nucleus, and the arsis (derived from the Greek ”to raise” the foot) represents the transient part between two nuclei. The coda of one syllable and the attack of the next one (if they exist) are grouped to form the arsis. If there are no coda and no attack, the arsis still exists and corresponds to a short transition between two vowels. Table 1 shows the syllables and the arsis and thesis splits of the word ”manual”. This word is made of three syllables. It contains three thesis, but four arsis. Controlling syllabic rhythm induces controlling these seven time points.

Table 1: *Example of a word (1), along with its phonetic transcription (2), and its split into syllables (3) and arsis and thesis (4) (arsis are inside brackets)*

(1)	Manual
(2)	_ m æ n j u ə l _
(3)	[_ m æ n] [j u] [ə l _]
(4)	[_ m] æ [n j] u [] ə [l _]

The *Syllabic Control Points* (SCP) are defined as temporal marks for rhythm control. An example of SCP for the sentence ”My name is” (/majnejmɪz/) is displayed in Figure 2. *Vocalic Points* (P_v) are the SCP that correspond to thesis in the vocalic nuclei , and *Transient Points* (P_t) those that correspond to arsis in the transient phases . These points define a target temporal location for each phase: when a vocalic phase is triggered (see section 2.4), the target time-instant aims at the corresponding P_v until the next transient phase is triggered. Once this transient phase is triggered, the target time-instant evolves from the current P_v to the next P_t, and the synthesis signal duplicates the original period aimed by this P_t until the next vocalic phase is triggered, and so on. On Figure 2, seven SCP are plotted on the spectrogram for the displayed sentence. Controlling the timing of these points allows for syllabic rhythm control while preserving the correct articulation.

Placement and position of the two SCP per syllable is an important issue. The first SCP is a P_t and the second one is a P_v. An example is shown on Figure 2. The P_v are placed in the center of the corresponding vowel to ensure its correct pronunciation. The P_t are placed around the center of the final consonant of a cluster (that may contain several phones) or in the center of a transition between two vowels. This guaranties an accurate control of the moment of occurrence of the next P-center when the next vocalic phase is triggered. If the last consonant of a cluster is an unvoiced plosive, the corresponding P_t should be placed during the silence prior the explosion. There is a special treatment for the first and the final P_t: to make sure that every phoneme is pronounced entirely, the first P_t should be placed at the end of the silence prior the first phoneme, and the final P_t should be placed at the beginning of the silence following the last phoneme.

2.4. Binary rhythm control: Tap mode

In *Tap mode*, arsis and thesis are controlled by tapping a control button, as shown on Figure 2. Pressing the control button triggers a vocalic phase (thesis), while releasing it triggers a transient phase (arsis). At the beginning, the first P_t is selected (i.e. $t_i = P_t(1)$). When the control button is pressed, the target time-instant evolves from the first P_t to the first P_v. Once this P_v is reached ($t_i = P_v(1)$), the corresponding pitch period or signal portion is repeated until the control button is released. Then, the target time-instant evolves from the current P_v ($P_v(1)$) to the next P_t ($P_t(2)$). If the control button is pressed again, the target time-instant evolves from the current

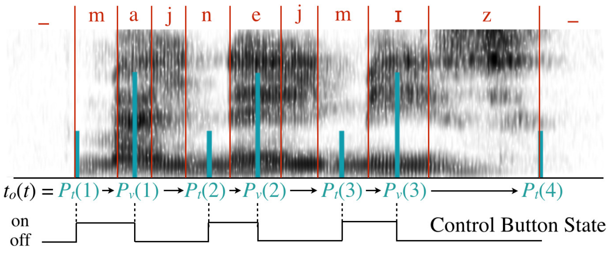
_
m
a
j
n
e
j
m
ɪ
z
_
$t_o(t) = P_t(1) \rightarrow P_v(1) \rightarrow P_t(2) \rightarrow P_v(2) \rightarrow P_t(3) \rightarrow P_v(3) \rightarrow P_t(4)$
on
off
Control Button State

Figure 2: *Syllabic Control Points for the sentence* ”my name is”. *From top to bottom: phonetic labels, spectrogram, syllable control points, syllabic rhythm control with a button.*

P_t ($P_t(2)$) to the next P_v ($P_v(2)$), and so on until the end of the original signal is reached. One syllable is pronounced by two motions, i.e. a release-pressure-release sequence. Note that transitions between two SCP are played at a predefined rate. This rate can be set independently for vowels, consonants and silences, but it can not be controlled in real time. Then in principle shortening of a recorded utterance can degrade the sound quality, as part of the signal may be truncated.

2.5. Continuous articulation control: Fader mode

The *Fader mode* allows for more accurate syllabic transitions control than the tap mode: the target time-instant during transient phases can be varied continuously using a *fader* controller, like e.g. an expression pedal. This control mode appeared very effective for singing synthesis articulation control, where syllabic rhythm is often much slower than in speech. It will not be further discussed in this paper.

3. Experiment in prosodic mimicry

3.1. Protocol

The prosodic control ability of a group of subjects using VOKinesiS has been formally assessed using a prosodic imitation paradigm. The subjects' task was to reproduce as accurately as possible the prosody of sentences, by drawing the melody on a Wacom Intuos 5 graphic tablet and tapping the rhythm with an iMac G5 keyboard space bar. Output pitch was set according to the stylus position on the tablet, and syllabic rhythm was set according to the space bar state, using SCP (see section 2.4).

A set of 8 sentences ranging from 2 to 9 syllables (the same as in [1]), recorded by a male and a female speaker, was presented in a random order to 8 subjects. Subjects could make as many trials as they wished. When they thought their performance was good enough, they could save the reproduced sentence and start recording the next one. The whole procedure had to be performed twice: the first time without pitch control, and the second time with pitch control. On average, the whole test lasted about 50 minutes. It was not particularly difficult, and the subjects were given only minimal training.

3.2. Measurement procedure

Original phonemes were labelled with Praat [15], using waveform and spectrogram visualisations. As in [16], any question about a particular boundary was resolved by listening to the segments around it, and the end of a sentence was determined by the end of periodicity.

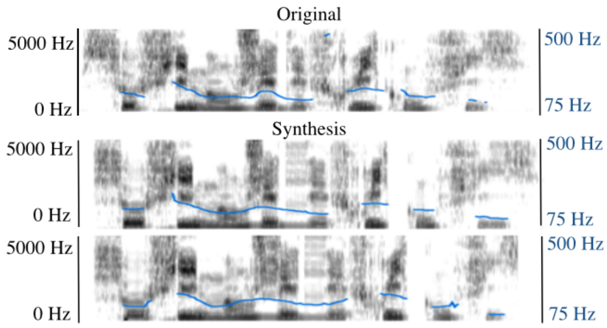

Figure 3: Original (top) and synthesis (middle and bottom) spectrograms and pitch contours for the sentence composed of 8 syllables. Frequency range is displayed on the left axis for spectrograms,and on the right axis for pitch contours.

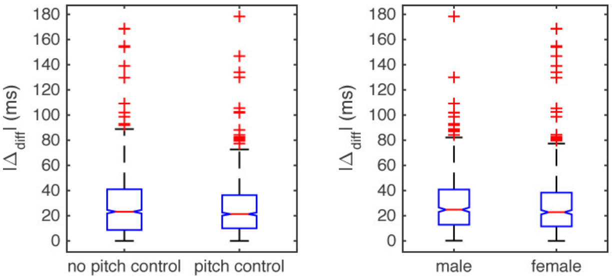

Figure 4: $|\Delta_{diff}|$ obtained with and without pitch control (left), for the male and the female speakers (right)

The position of the target time-instant was recorded during each performance. This position was then used to determine the synthesis phoneme labels, according to the original labels. Durations of Inter-Vocalic (I-V) units (i.e. durations between vowel onsets) have been measured out of phoneme labels as follows

$$\Delta(s) = v_b(s+1) - v_b(s) \qquad (1)$$

where s is the syllable index and $v_b(s)$ the beginning of the corresponding vowel. I-V durations are noted Δ_{in} for the original signals and Δ_{out} for the re-synthesized signals. Differences between Δ_{in} and Δ_{out} have been used to assess rhythm reproduction accuracy:

$$\Delta_{diff}(s) = \Delta_{in}(s) - \Delta_{out}(s) \qquad (2)$$

3.3. Results

Pitch control accuracy using a graphic tablet has already been assessed for intonational and melodic control [1, 9], and will only be briefly discussed. This section will focus on syllabic rhythm control accuracy.

3.3.1. Pitch control

Figure 3 displays pitch contours of the sentence composed of 8 syllables, recorded by the male speaker, for natural speech and for 2 gestural reproductions. Although pitch contours evolve with less abrupt changes for synthesis, the overall shape of the original pitch contour is conserved, as was expected according to [1].

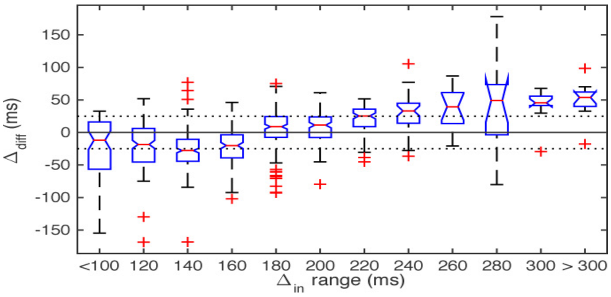

Figure 5: Δ_{diff} for different Δ_{in} ranges, varying from less than $100ms$ to more than $300ms$.

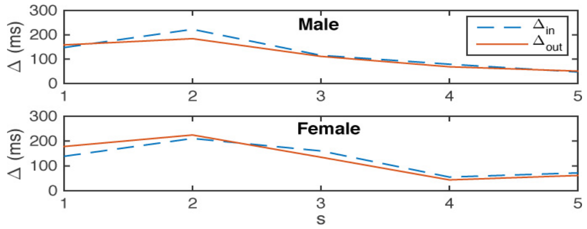

Figure 6: Δ_{in} (dashed lines) and mean(Δ_{out}) (plain lines) for the sentence composed of 5 syllables, spoken by the male (top) and the female (bottom) speakers, for all subjects, with and without pitch control.

3.3.2. Rhythm control

On average, for all subjects and sentences, $|\Delta_{diff}|$ is about 20 ms, which corresponds to two periods at 100Hz. One can conclude that rhythmic reproduction can be performed with excellent accuracy. Figure 4 shows that neither pitch control nor original speaker have significant effect on rhythm reproduction accuracy: $mean(|\Delta_{diff}|)$ is always around $20ms$. Figure 5 shows two effects: 1) long syllables ($> 220ms$) are more difficult to reproduce, since $mean(|\Delta_{diff}|)$ can reach $50ms$, whereas for shorter syllables, $mean(|\Delta_{diff}|)$ has a maximal value of $25ms$; 2) a tendency to shorten long syllables and to lengthen short syllables. However, Figure 6 shows that subjects were able to follow the original I-V duration variability within a sentence.

4. Discussion

VOKinesiS allows for accurate rhythm control, by focusing on syllabic frame timing control. More accuracy in articulation timing can be obtained using faders instead of a button. The principles for syllabic rhythm control are discussed in this section, as well as the possible applications in speech research.

4.1. Syllabic Frame Rhythm Control

The concept of SCP for syllabic frame rhythm control is not in contradiction with the concept of P-centers for syllabic rhythm perception: placing a P_t close to the next vowel onset allows to control the moment of occurrence of the P-center when a vocalic part is triggered. Despite the fact that SCP are used for syllabic rhythm control, this concept can also be used for stress and mora timed languages : any language is built on a syllabic basis [2, 10, 11]. For e.g. stress timed languages, a full control of rhythm consists in controlling stresses with pitch variations (i.e. with the stylus) along with the underlying syllabic rhythmic patterns (i.e. with the button). For musical purposes, the concept of SCP can be extended, beyond voice sounds, to any sound sample enriched with labels that can be used as SCP.

Although the work presented here only focused on binary rhythm control with a button, VOKinesiS can be used with any continuous controller, assigned to any limb. Another part of our future work will consist in assessing rhythm control with a continuous foot pedal for speech and singing synthesis, and to evaluate which control method is more suited for speech or singing.

4.2. Applications

VOKinesiS is primarily a performative speech and singing instrument: it has been designed with musical applications and expressive speech synthesis in mind (see Additional material) [8, 17]. Performative rhythm control is also a new paradigm that can be useful for prosodic research. Although finger tapping experiments have been used to determine the location of the P-centers in syllables [18, 19], the release movement has never been studied. Measurements of release variations during tapping tasks to determine whether P-centers consist in only one or two temporal parameters would be interesting to explore. VOKinesiS could be used to compare tapping movements performed 1) by following isochrone reiterant speech recorded along with a metronome, 2) by producing reiterant speech with VOKinesiS along with a metronome.

Brain imaging studies of the use of VOKinesiS could be very useful to verify or to provide new insights to some aspects of brain organisation theories about voice production and motor planning, like e.g. [2, 20, 21]. VOKinesiS could provide new speech therapy techniques. It could help patients who suffer from Broca's aphasia [22] to recover language/facilitate communication: the fact that the Melodic Intonation Therapy (MIT) is one of the most effective therapies seems to be due, to a large extent, to the fact that finger tapping is performed during syllabic production tasks [23]. Furthermore, some efficient therapeutic techniques to treat stuttering seem to "synchronize a disturbed signal transmission between auditory, speech motor planning, and motor areas" [24]. Thus, the system might also be used to help people who stutter to learn how to produce fluent speech by e.g. reading a text while controlling the same pre-recorded text with VOKinesiS.

4.3. Additional material: sound examples

The attached folder provides examples of rhythmic patterns produced with VOKinesiS. The file *BD9.wav* contains the original sentence composed of 9 syllables recorded by the male speaker. Gestural reproductions with and without pitch control (*BD9pitch* and *BD9nopitch*) are provided. Musical improvisations recorded by an amateur musician can be heard in the files *BD9song.wav*, *BMmusic.wav* and *BMnomus.wav* (the two latter contain the exact same modification of *BM.wav*, with and without musical background). Even if some notes are out of tune, musical syllabic rhythm is controlled with reasonable accuracy, for both French (*BD9*) and English (*BM*).

5. Acknowledgements

This work was supported by the Agence Nationale de la Recherche ANR, through the ChaNTeR project (ANR-13-CORD-0011, 2014-2017).

6. References

[1] C. d'Alessandro, A. Rilliard, and S. Le Beux, "Chironomic stylization of intonation," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1594–1604, 2011.

[2] P. F. MacNeilage, "The frame/content theory of evolution of speech production," *Behavioral and brain sciences*, vol. 21, no. 04, pp. 499–511, 1998.

[3] M. Astrinaki, "Peformative statistical parametric speech synthesis applied to interactive designs," Ph.D. dissertation, University of Mons, 2014.

[4] N. D'Alessandro and T. Dutoit, "Handsketch bi-manual controller: investigation on expressive control issues of an augmented tablet," in *Proceedings of the 7th international conference on New interfaces for musical expression*. ACM, 2007, pp. 78–81.

[5] L. Feugère, C. d'Alessandro, B. Doval, and O. Perrotin, "Cantor digitalis: chironomic parametric synthesis of singing," *EURASIP Journal on Audio, Speech, and Music*, 2017.

[6] S. Le Beux, B. Doval, and C. d'Alessandro, "Issues and solutions related to real-time td-psola implementation," in *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[7] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech communication*, vol. 16, no. 2, pp. 175–205, 1995.

[8] M. Evrard, S. Delalez, C. d'Alessandro, and A. Rilliard, "Comparison of chironomic stylization versus statistical modeling of prosody for expressive speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] C. d'Alessandro, L. Feugere, S. Le Beux, O. Perrotin, and A. Rilliard, "Drawing melodies: Evaluation of chironomic singing synthesis," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3601–3612, 2014.

[10] P. Mairano, "Rhythm typology: acoustic and perceptive studies," Ph.D. dissertation, Università degli studi di Torino, 2011.

[11] B. r. Lindblom, "Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns," *Phonetica*, vol. 57, no. 2-4, pp. 297–314, 2000.

[12] S. K. Scott, "Perceptual centers in speech - an acoustic analysis," Ph.D. dissertation, University College London (University of London), 1993.

[13] P. A. Barbosa, P. Arantes, A. R. Meireles, and J. M. Vieira, "Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors." in *Interspeech*, 2005, pp. 1441–1444.

[14] P. Wagner, "The rhythm of language and speech: Constraining factors, models, metrics and applications," Ph.D. dissertation, Habilitationsschrift, University of Bonn, 2008.

[15] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[16] A. G. Levitt, "Reiterant speech as a test of non-native speakers' mastery of the timing of french," *The Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 3008–3018, 1991.

[17] S. Delalez and C. d'Alessandro, "Vokinesis: syllabic control points for performative singing synthesis," in *NIME*, vol. 17, 2017.

[18] B. H. Repp, "Sensorimotor synchronization: a review of the tapping literature," *Psychonomic bulletin & review*, vol. 12, no. 6, pp. 969–992, 2005.

[19] R. C. Villing, B. H. Repp, T. E. Ward, and J. M. Timoney, "Measuring perceptual centers using the phase correction response," *Attention, Perception, & Psychophysics*, vol. 73, no. 5, pp. 1614–1629, 2011.

[20] J. Houde, S. Nagarajan, and M. Merzenich, "Modulation of auditory cortex during speech production: An meg study," in *Proceedings of the Fifth Seminar on Speech Production: Models and Data.(Munich, Germany: Unversitat Munchen)*, 2000, pp. 249–252.

[21] Z. Z. Zheng, K. G. Munhall, and I. S. Johnsrude, "Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production," *Journal of cognitive neuroscience*, vol. 22, no. 8, pp. 1770–1781, 2010.

[22] P. Broca, "Deux cas d'aphemie tramatique, produite par des lesions de la troisieme circonvolution frontale gauche," *Bulletin de la Societe de Chirurgie de Paris, V*, pp. 51–54, 1864.

[23] G. Schlaug, S. Marchina, and A. Norton, "From singing to speaking: why singing may lead to recovery of expressive language function in patients with broca's aphasia," *Music perception: An interdisciplinary journal*, vol. 25, no. 4, pp. 315–323, 2008.

[24] K. Neumann, H. A. Euler, A. W. von Gudenberg, A.-L. Giraud, H. Lanfermann, V. Gall, and C. Preibisch, "The nature and treatment of stuttering as revealed by fmri: A within-and between-group comparison," *Journal of fluency disorders*, vol. 28, no. 4, pp. 381–410, 2004.