Assessment of finger tapping for rhythm control in performative speech synthesis

Christophe d'Alessandro, Grégoire Locqueville

Institut Jean le Rond ∂'Alembert, Sorbonne Université - CNRS, Paris (FRANCE)
christophe.dalessandro@sorbonne-universite.fr, gregoire.locqueville@ sorbonne-universite.fr

Performative Vocal Synthesis (PVS), developed initially as a new type of musical instrument, allows the real-time gestural control of synthesized speech through the modulation of voice pitch, syllable timing, vocal effort, and vocal quality [1,2]. As an interaction technique, PVS gives a "prothetic" voice whose gestures for control are explicit and externalized, contrary to the natural voice. It has been used for computer-aided intonation training in second language learning [3] and voice and motor reeducation (see the Gepeto project http://gepeto.dalembert.upmc.fr/).

The present research aims to assess the precision of finger taping for rhythm production, as it is used in the PVS system Voks [4]. The taping rhythm control paradigm is based on an analogy with speech production's frame/content theory [5]. Syllables are considered time "frames" (cycles of articulators open-close alternation) where segmental "contents" (phonemes) take place. These opening and closing cycles can be exploited for rhythmic control with the help of Syllabic Control Points (SCP). Vocalic Points (Pv) correspond to vocalic nuclei, and Intervocalic Points (Pi) correspond to intervocalic consonants (syllabic attack and coda). When the finger depresses the button, a Pv is triggered; when the finger raises the button, the Pi is triggered. Figure 1 shows the placement of Pv and Pi on the signal and the process of rhythm control using tapping.

The precision achieved by finger taping for rhythmic control is assessed using a prosodic imitation paradigm and a prosodic synchronization paradigm. A set of 8 French sentences ranging from 2 to 9 syllables, recorded by a male and a female speaker, is presented randomly to 8 subjects. In the imitation task, subjects listen to a sentence and are asked to reproduce it using Voks. The subjects' task is to reproduce as accurately as possible the prosody of sentences after listening to them, tapping the rhythm with a MacBook keyboard space bar (a non-synchronized motor repetition task [6]). In the synchronization task, they are asked to play the synthetic and natural sentence in synchrony (sensorimotor synchronization task [6]). The question whether biphasic control points (using Pv and Pi) or monophasic control points (using Pv or P-center only) are needed is investigated: sensorimotor synchronization experiments assume that each tap aims at only one rhythmic anchor [7] although motor control gestures are intrinsically biphasic (lowering and rising the finger, opening and closing the vocal tract).

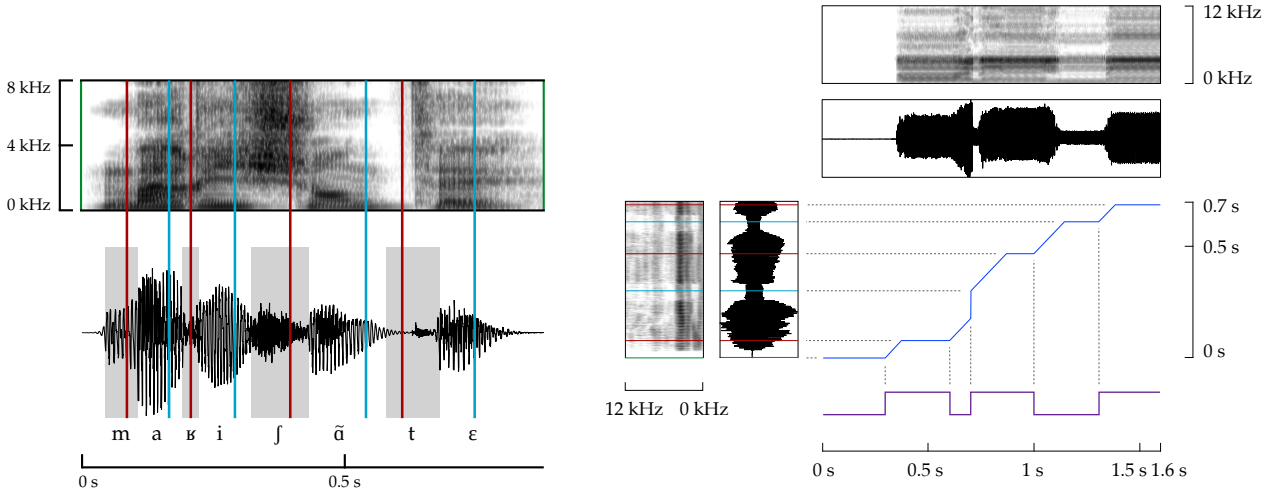

Figure 1. (Left) French sentence *Marie chantait*. Vocalic control points are in red; intervocalic control points are in cyan. (Right) Principle of biphasic taping rhythm control. Y-axis: original sound. Y-axis: tap sequences (bottom and synthesized sound. The time index is displayed as the synthesis time as a function of the original timeline.

Rhythmic precision is measured in this preliminary experiment for :1. comparing vocal and tapping rhythmic precision; 2. studying the effect of dominant/non-dominant hand; 3. studying the effect one or two fingers tapping; 4. studying the effect of simultaneous rhythmic and pitch control; 5. studying the effect of one vs two SCP per syllable. The tested conditions are: **Nat.** natural vocal production; **Cont. Pts**: monophasic tasks (one point per syllable, using Pv); **P-cent**. Monophasic tasks (one point per syllable, using P-centers); **Base**: biphasic task using Pv and Pi; **Hnd 2**: using the subjects' nondominant hand; **2 Fngs**: Using two fingers; **Tblt**: Simultaneously controlling pitch (as a distractor to the rhythmic task). The mean distance between production and stimulus phoneme boundaries is used to assess the precision obtained for each condition. Results are reported in Figure 2 (left: imitation, right: synchronization).

This preliminary experiment suggests that: 1. subjects are better at reproducing and synchronizing rhythm with their natural voice than with tapping; 2. the best tapping condition is biphasic with one finger; 3. one finger performs better than two; 4. the preferred hand performs slightly better; 5. simultaneous pitch and rhythm control is impairing precision; 5. monophasic conditions with P-centers and Pv are comparable; 6. the synchronization task is slightly less precise than the imitation task; 7. Biphasic control il more precise than Monophasic control.

In summary, the biphasic tapping paradigm shows good precision for performative rhythm control in speech. The experiments support the hypothesis that tapping using biphasic control points is a motor control process somewhat analogous to articulatory oscillation for syllable production in speech. Further studies are needed to extend this preliminary study to more subjects and to other language. Anticipation and of taps and individual strategies in rhythmic control must also be investigated.

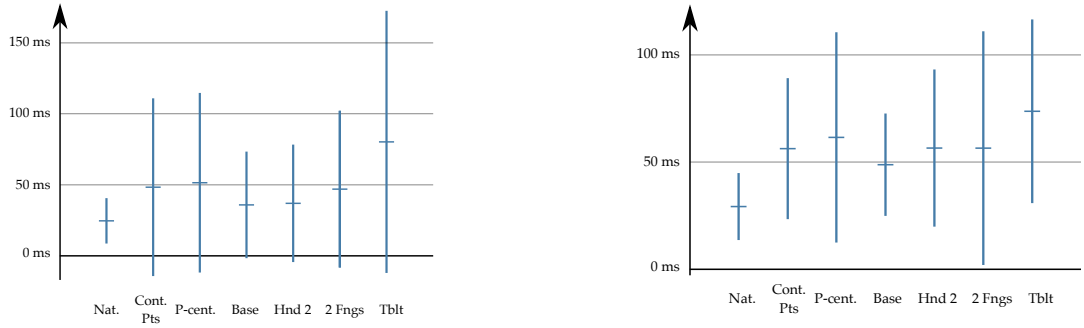

Figure 2. Mean and standard deviation of the absolute shift for imitation (left) and synchronization (right) tasks for all sentences with all subjects.

[1] Christophe d'Alessandro, Albert Rilliard, and Sylvain LeBeux (2011). Chironomic stylization of intonation. Journal of the Acoustical Society of America, 129(3):1594-1604, 2011.

[2] Samuel Delalez and Christophe d'Alessandro (2017). Adjusting the Frame: Biphasic Performative Control of Speech Rhythm. Proc. Interspeech 2017, 864-868, Stockholm, Sweden.

[3] F Xiao Xiao, Nicolas Audibert, Grégoire Locqueville, Christophe d'Alessandro, Barbara Kuhnert, and Claire Pillot-Loiseau (2021). Prosodic disambiguation using chironomic stylization of intonation with native and non-native speakers. Proc. Interspeech 2021,516-520. Brno, Czech Republic.

[4] Grégoire Locqueville, Christophe d'Alessandro, Samuel Delalez, Boris Doval, and Xiao Xiao (2020). Voks: Digital instruments for chironomic control of voice samples. Speech Communication, 125:97-113.

[5] Peter F. MacNeilage (1998). The frame/content theory of evolution of speech production. Behavioral and Brain Sciences, 21(4):499-511.

[6] Ho Tamara Rathcke, Chia-Yuan Lin, Simone Falk and Simone Dalla Bella (2021), Tapping into linguistic rhythm. Laboratory Phonology: Journal of the Association for Laboratory Phonology 12(1): 11, pp. 1–32, 2021.

[7] Bruno H. Repp (2005). Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin & Review, 12(6):969-992.