WHEN LANGUAGE HITS THE BEAT: SYNCHRONISING MOVEMENT TO SIMPLE TONAL AND VERBAL STIMULI

Tamara Rathcke[1], Chia-Yuan Lin[1], Simone Falk[2,3], Simone Dalla Bella[3]

[1]University of Kent, [2]University Sorbonne Nouvelle Paris-3, [3]University of Montreal
t.v.rathcke@kent.ac.uk; c.lin@kent.ac.uk; simone.falk@univ-paris3.fr; simone.dalla.bella@umontreal.ca

ABSTRACT

Rhythmic perception-action coupling through sensorimotor synchronization has been studied with non-verbal, simple and complex auditory signals like a metronome and music. Applications of the paradigm to language are relatively rare, but could provide a valuable tool for investigating rhythm perception in speech. The aim of the present study is to compare sensorimotor synchronisation with simple non-verbal and verbal stimuli.

Twenty-nine English-speaking participants tapped in synchrony with, and after listening to, a set of pure tones and simple syllables at three different tempi. Synchronisation to the vowel onset of verbal stimuli was comparable to the synchronisation to the acoustic onset of simple tones. Stability of inter-tap intervals decreased in the non-synchronised continuation condition at a slower tempo. These findings suggest that similar perceptual mechanisms are in place for simple auditory stimuli, regardless of their origin and complexity, and support the idea that processing and encoding of linguistic prosody relies on general aspects of the perceptual and motor system.

Keywords: speech rhythm; rhythm perception; sensorimotor synchronisation; movement.

1. INTRODUCTION

Rhythm perception in music is sometimes viewed as based on the extraction of a beat – a steady, repeating (and therefore predictable), subjectively prominent pulse that can be tracked even among temporal irregularities of a complex signal [9]. The perceived beat is known to evoke time-locked body movements in listeners, most readily observable in dance or other forms of embodied responses to music. In healthy individuals, such beat perception and synchronisation do not require much conscious effort. Moving to the beat of music is a natural and widespread response to a subjectively experienced beat.

A growing body of research has exploited this ability of external, rhythmically structured events to entrain movement. The sensorimotor synchronisation (SMS) paradigm has been developed and successfully utilised as a laboratory tool to study rhythm perception and production, as well as the properties of the human timing system by observing how a motor action (e.g. finger tapping) is temporally coordinated with an external auditory event [2, 17].

In contrast, the nature of linguistic rhythm has been a matter of controversial debates [3, 19]. Neither speech production research [1, 6], nor perception experiments [13, 15] have helped to resolve the controversy. The present study aims to test a new movement-based paradigm to address the issue.

SMS with language has been studied only rarely [e.g. 8, 11] as it is generally agreed that in contrast to music, language does not naturally entrain movement [4]. Moreover, SMS with language displays a high amount of variation in comparison to SMS with music, cf. coefficients of variation of 30% vs. 4%, respectively (after [4]). Yet despite the issue of an increased variability, SMS seems to capture some core properties of rhythmic variation both between languages [11] and within a language [8].

The SMS performance might appear poorer with language than with music because acoustic targets of synchronisation are more difficult to define. For this reason, previous studies of SMS with language [8, 11] avoided to apply any common measures of synchronisation accuracy [2] and calculated measures of tapping variability instead. Alternatively, this methodological issue might be alleviated by using a non-synchronised continuation (NSC) paradigm where listeners are asked to tap the perceived rhythm after they had listened to an auditory prompt [20].

The present study aimed: (1) to provide evidence on the lowest limit of SMS variability with language; (2) to identify the most likely acoustic anchors for SMS in verbal stimuli; (3) to compare SMS to NSC, in order to guide methodological decisions on the choice of an appropriate movement paradigm in future research.

2. METHOD

2.1. Stimuli

The overall experiment included different types of stimuli, though this paper focuses on the comparison between (1) pure tones and (2) simple syllables.

The set of pure-tone stimuli contained high (H, 260 Hz) and low (L, 130 Hz) tones, created in Praat. The verbal set contained monosyllables [bi] and [bu]

(reminiscent of real English words *bee* and *boo*), produced by a native male speaker of Greek with fully voiced stop closures. The pitch was normalised to 130 Hz with a slight declination slope. Both syllables and tones were 250 ms in duration, each followed by a 50-ms silence. They were then combined into sequences with three inter-onset intervals (IOI) between targets: 300, 600 and 1200 ms (see Table 1). The created 12 stimuli in three different tempi of target occurrence: fast (300 ms IOI), intermediate (600 ms IOI) and slow (1200 ms IOI). These stimuli were repeated 20 times in SMS tasks and 10 times in NSC tasks.

Table 1: Outline of the material design.

Stimulus	Target	IOI (ms)		
		300	600	1200
Tonal	L H	L H	L H H L	L H H H H L L L
Verbal	bi bu	bi bu	bi bu bu bi	bi bu bu bu bu bi bi bi

2.2. Participants

Twenty-nine native English-speaking participants (8 male, mean age: 23 years) took part in this research. They self-reported no known history of speech, writing or hearing problems, and no motor disorders.

2.3. Tasks and procedure

Prior to the SMS experiment, participants were asked to fill in an online questionnaire that ran the relevant health checks (see 2.2) and collected some demographic information.

Participants were then asked to perform a selection of tests from the *Battery for the Assessment of Auditory Sensorimotor Timing Abilities* (BAASTA) [5]. The tests included (1) tapping to a metronome at 450 and 600 ms IOI, and (2) self-paced tapping at the individually most comfortable speed and at the fastest possible speed.

During the main phase of the experiment, all stimuli were tested with two tasks: synchronisation (SMS) and continuation (NSC). When synchronising, participants were presented with 20 repetitions of each stimulus and asked to tap the index finger of their dominant hand in time with the prompted target [2] (see Table 1). When continuing a rhythmic pattern, participants were requested to listen silently to 10 repetitions of the stimulus first and then tap its rhythm once the auditory playback had stopped. The task order was counterbalanced using the Latin-square design. All stimuli were presented in a random order.

The data were collected on a Roland HandSonic drum pad and a Dell Latitude 7390 laptop. The overall duration of the experiment varied across individual session but overall, it was no longer than 45 minutes.

2.4. Analyses

A set of measures was calculated to describe the degree of synchrony between produced taps and acoustic targets [2, 17]. Acoustic targets under scrutiny here were the local maximum amplitude and the stimulus onset in case of pure tones, or the maximum amplitude, the syllable and vowel onset in case of verbal stimuli.[1]

We ran mixed-model statistics due to an imbalanced dataset (participant was the only random intercept). The following measures of SMS performance were tested as dependent variables:

- **Absolute asynchronies** (AA, see 3.1): the models were fit for the tonal and the verbal stimuli separately. For pure tones, predictors included *acoustic target* (onset vs. maximum amplitude), *prompted target* (H vs. L), *IOI* (300, 600, 1200 ms) and individual AA performance with the metronomes (mean AA at 450 and 600 ms). For verbal stimuli, the fixed effect structure was similar, with the notable differences of *acoustic target* (which had 3 levels – syllable onset,[2] vowel onset, timestamp of the maximum amplitude) and *prompted target* (which had the two levels [bi] and [bu]). We only tested for one interaction – *acoustic* and *prompted target* – to check if synchronisation anchors may change for different prompts.
- **Signed asynchronies** (SA, see 3.2): as above, the tonal and the spoken stimuli were treated in two separate models due to their differences in the plausibly assumed acoustic targets of synchronisation. But in contrast to AA models, participant's SA with the metronomes was used as the measure of individual SMS performance.
- **SMS consistency** (see 3.3): the dependent variable here was the standard error of asynchrony (SE). An effect structure similar to the above (AA/SA) was adopted, though the measure of individual SMS ability was participant's SE with the metronomes.

NSC was evaluated by measuring the variability of inter-tap intervals (ITI) and comparing it to SMS. For this, we calculated the coefficient of variation CV of the ITIs, following the formula in (1) (see 3.4), and fit a model with the fixed predictors *stimulus type* (tonal vs. verbal), *task* (SMS vs. NSC), *IOI* (300, 600, 1200 ms) and the individual variability of spontaneous tapping in self-paced and fast conditions. In this model, we tested for all possible 2-way interactions.

$$(1) \quad CV(ITI) = \left(\frac{SD(ITI)}{mean(ITI)}\right) \times 100$$

3. RESULTS

3.1. Absolute asynchronies

AA shows the synchronisation accuracy (in % of the target IOI) for an acoustic target and a produced tap. Larger AA indicates lower SMS accuracy.

3.1.1. Tonal stimuli

The best-fit model for AA with tones showed an effect of *IOI* ($F(2)=61.8$, $p<0.001$), *acoustic target* ($F(1)=163.4$, $p<0.001$) and *prompted target* ($F(1)=25.3$, $p<0.001$). Accordingly, AA was larger for 300 than 600 ms IOI (14.0% vs. 9.3%, $t=3.8$, $p<0.001$) and again larger for 600 than 1200 ms (9.3% vs. 0.5%, $t=7.2$, $p<0.001$). The model further identified that the intensity maximum was a poor synchronisation anchor in these stimuli: the accuracy improved by approximately 12.6% when measured with respect to an acoustic stimulus onset in comparison to a local amplitude maximum ($t=12.8$, $p<0.001$). Participants' synchrony with the stimulus onset was slightly better in L than in H tone sequences (4.4% vs. 9.3%, $t=5.0$, $p<0.001$).

Moreover, participants who tapped with larger asynchronies to a fast-paced metronome (450 ms IOI), also had larger AA values with the tonal stimuli ($F(1)=17.9$, $p<0.001$). In contrast, their performance with the slow-paced metronome (600 ms IOI) did not matter for this task.

3.1.2. Verbal stimuli

The best-fit model for AA with speech showed a main effect of *IOI* ($F(2)=73.0$, $p<0.001$) and an interaction of *acoustic* and *prompted target* ($F(2)=9.7$, $p<0.001$). In keeping with the results for the tonal stimuli, AA was larger for 300 than 600 ms IOI (8.8% vs. 6.1%, $t=4.8$, $p<0.001$) and again larger for 600 than 1200 s (6.1% vs. 2.0%, $t=7.2$, $p<0.001$). The interaction essentially demonstrated that the location of the intensity maximum served as a poor tap attractor for [bu] but not [bi] (9.5% vs. 5.7%, $t=4.8$, $p<0.001$). For [bi], the *syllable* onset was a slightly better acoustic target than the *vowel* onset (3.9% vs. 6.1%). A similar trend was found for [bu], but the effect did not reach the set significance level (5.4% vs. 3.9%, $t=1.8$, $p=0.065$). Once again, participants' tapping performance with a fast-paced metronome was indicative of their performance with the verbal stimuli ($F(1)=18.1$, $p<0.001$).

3.1.3. Tonal and verbal stimuli compared

To compare SMS of tonal and spoken stimuli in a single model, we removed the intensity maximum as the synchronisation target and ran two models, looking for a converging performance between SMS to pure tones and SMS to spoken stimuli (with either the syllable or the vowel onset as the acoustic target of synchronisation in that latter case). The difference between AA of tonal and verbal stimuli was only significant for syllable ($F(1)=4.8$, $p<0.05$) but not vowel onsets.

3.2. Signed asynchronies

SA shows if a tap preceded or followed an acoustic target of synchronisation. A negative value (in % of the target IOI) indicates that the tap anticipated the synchronisation target.

3.2.1. Tonal stimuli

The best-fit model for SA with tones included *IOI* ($F(2)=61.8$, $p<0.001$) and *acoustic target* ($F(1)=163.4$, $p<0.001$) as the only significant effects. Accordingly, taps tended to precede acoustic targets at the shortest IOI of 300 ms (-7.1%) but followed the target at longer IOIs of 600 or 1200 ms (3-5%, $t>5.8$, $p<0.001$). If the acoustic target was defined by an intensity maximum, taps showed a relatively large negative asynchrony (-15.3%, $t=11.0$, $p<0.001$).

3.2.2. Verbal stimuli

The best-fit model for SA with verbal stimuli included *IOI* ($F(2)=23.8$, $p<0.001$) and an interaction of *acoustic* and *prompted target* ($F(2)=3.6$, $p<0.05$). In contrast to SMS with tones, SMS with speech differed significantly at each IOI. Measured at the vowel onset, the delay was -4.3% at 300 ms vs. 5.6% at 600 ms ($t=6.9$, $p<0.001$) vs. -1.7% at 1200 ms ($t=3.9$, $p<0.001$).

Once again, we found differences between SMS with [bi] vs. [bu] that only arose for the intensity maximum as the potential synchronisation target, with [bu] having a negative SA of -16.1% and [bi] a negative SA of -2.0%, $t=3.3$, $p<0.001$).

3.2.3. Tonal and verbal stimuli compared

Again, we compared synchronisation with tonal and spoken stimuli in a single model (see 3.1.3). Accordingly, tonal and verbal stimuli differed significantly if the syllable onset was considered the target ($F(1)=11.1$, $p<0.001$) but not if the vowel onset was the synchronisation anchor.

3.3. SMS consistency

SE (again in % of IOI) captures the SMS consistency by examining the standard error of asynchronies. Larger values indicate poorer SMS.

3.3.1. Tonal stimuli

The best-fit model for SE with tones included *IOI* ($F(2)=13.8$, $p<0.001$) and *acoustic target* ($F(1)=5.3$, $p<0.05$). Tapping variability was higher at slower tempi (300/600 ms vs. 1200 ms, $t>4.4$, $p<0.001$). If the intensity maximum was chosen as the SMS anchor, SE decreased by 0.4% ($t=2.3$, $p<0.05$). Finally, SE with both metronomes was predictive of the participants' SE with tones (450 ms IOI: $F(1)=7.4$, $p<0.05$; 600 ms IOI: $F(1)=5.5$, $p<0.05$).

3.3.2. Verbal stimuli

The best-fit model for SE with verbal stimuli had *IOI* ($F(2)=98.3$, $p<0.001$) and *acoustic target* ($F(2)=3.6$, $p<0.05$). Accordingly, tap variability decreased at longer IOIs (300/600 ms: 1.5/1.1%, $t=5.7$, $p<0.001$; 600/1200 ms: 1.1/0.5%, $t=8.1$, $p<0.001$). SMS with the vowel onset increased SE, as compared to the syllable onset or the intensity maximum (with identical results for both comparisons, 1.3% vs. 1.1%, $t=2.3$, $p<0.05$). Participants who tapped more variably to a slow-paced metronome (600 ms IOI), also had higher SE when tapping with verbal stimuli ($F(1)=8.6$, $p<0.01$).

3.1.3. Tonal and verbal stimuli compared

When SE of tonal and verbal stimuli were compared in a single model (similar to 3.1.3 and 3.2.3), no significant differences in variability were unveiled.

3.4. ITI consistency

The final measure captures the tapping consistency in SMS vs. NSC by examining the degree of variability across all inter-tap intervals (ITI). The higher the CV value, the less consistent the SMS/NSC performance.

The best-fit model produced an effect of *stimulus type* ($F(1)=8.2$, $p<0.01$) and an interaction of *IOI* and *task* ($F(2)=93.3$, $p<0.001$). Movement to tonal stimuli was slightly more variable than movement to spoken stimuli (4.3% vs. 4.0%, $t=2.9$, $p<0.01$). Figure 1 displays the interaction of *IOI* and *task*. At the shortest IOI, NSC showed slightly less variability than SMS (0.8%, $t=3.8$, $p<0.001$). However, SMS became increasingly less variable at larger IOIs while the opposite was true for NSC. The effect was smaller at 600 ms IOI (1%, $t=4.6$, $p<0.001$) and relatively large at 1200 ms IOI (3%, $t=15.3$, $p<0.001$).

4. DISCUSSION

The study demonstrated that SMS accuracy and variability with verbal prompts can be comparable to the performance obtained with tonal stimuli. Using a distractor paradigm for SMS [18], previous research has similarly shown that the discrepancy between music and speech in their ability to disturb SMS disappeared when they shared the same meter [4].

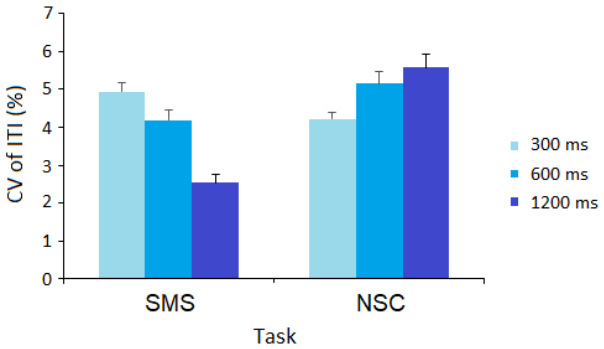
Unexpectedly, spoken stimuli could even help to slightly reduce the ITI variability in comparison to tonal stimuli. The effect might be due to a richer harmonic structure of speech as compared to pure tones, given that spectral discontinuities [21] and rise time of the amplitude envelope [10] are likely to be of critical importance for the perceptual extraction of a rhythmic event.

Furthermore, the results showed that the most likely acoustic anchors for SMS with verbal stimuli were the vowel (and not the syllable) onsets. In contrast, local intensity maxima served as poor SMS anchors in all stimuli, and they are also known to be poor predictors of the p-centre location [12]. Given voicing during the bilabial closure in our verbal stimuli, a tone onset and a syllable onset of speech were acoustically very similar. However, the vowel onset displayed the moment of the largest spectral discontinuity and was more comparable to the tone onset in terms of its ability to carry pitch.

Finally, our comparison of tapping consistency between SMS and NSC revealed that SMS was superior to NSC at slower tempi. SMS performance benefitted from the well-established subdivision effect [16] which NSC lacked. If, as we assume, the investigated IOIs may correspond to a syllable rate (300 ms), an inter-stress interval (600 ms) and a phrase-level interval (1200 ms) in real speech, NSC is likely to result in misleading conclusion about the most relevant, higher levels of rhythmic organisation.

The present study contributed to the understanding of beat perception in language using SMS-based paradigms. Our findings suggest that beat perception relies on a domain-general mechanism that can be engaged by verbal and tonal stimuli alike. These results support the idea that processing and encoding of linguistic prosody relies on general aspects of the perceptual and motor system [14].

Figure 1: Tapping consistency (means and standard errors of ITI) comparing SMS and NSC. The three IOIs are indicated in different shades of blue.



ACKNOWLEDGEMENTS

This research was supported by a research grant from the Leverhulme Trust (RPG-2017-306) to the first author.

4. REFERENCES

[1] Arvaniti, A., Rodriquez, T. 2013. The role of rhythm class, speaking rate and F0 in language discrimination. *Laboratory Phonology 4*, 7-38.

[2] Aschersleben, G. 2002. Temporal control of movements in sensorimotor synchronization. *Brain Cogn. 48*, 66–79.

[3] Cummins, F. 2012. Looking for rhythm in speech. *Empirical Musicology Review* 7(1-2), 28-35.

[4] Dalla Bella, S., Białuńska, A., Sowiński, J. 2013. Why movement is captured by music, but less by speech: Role of temporal regularity. *PLoS One 8(8),* e71945.

[5] Dalla Bella S. Farrugia, N., Benoit, C-E., Begel, V., Verga, L., Harding, E., Kotz, S. A. 2017. BAASTA: Battery for the Assessment of Auditory Sensorimotor Timing Abilities. *Behav. Res. Methods. 49,* 1128–1145.

[6] Dauer, R. M. 1983. Stress timing and syllable-timing reanalysed. *Journal of Phonetics 11,* 51-62.

[7] Drewing, K., Aschersleben, G., Li, S. C. 2006. Sensorimotor synchronization across the life span. *Int. J. Behav. Dev. 30*, 280–287.

[8] Falk, S., Rathcke, T., Dalla Bella, S. 2014. When Speech Sounds Like Music. *Journal of Experimental Psychology: Human Perception and Performance 40,* 1491-1506.

[9] Large, E., Palmer, C. 2002. Perceiving temporal regularity in music. *Cognitive Science 26(1),* 1-37.

[10] Leong, V., Hämäläinen, J., Soltész, F., Goswami, U. 2011. Rise time perception and detection of syllable stress in adults with developmental dyslexia. *Journal of Memory and Language 64*, 59-73.

[11] Lidji, P., Palmer, C., Peretz, I., Morningstar, M. 2011. Listeners feel the beat: Entrainment to English and French speech rhythms. *Psychon. Bull. Rev. 18,* 1035–1041.

[12] Marcus, S. M. 1981. Acoustic determinants of perceptual-center (P-center) location. *Perception and Psychophysics 30*, 247–256.

[13] Miller, M. 1984. On the perception of rhythm. *Journal of Phonetics 12,* 75-83.

[14] Parrell, B., Goldstein, L., Lee, S., Byrd, D. 2014. Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics 42,* 1-11.

[15] Rathcke, T., Smith, R. 2015. Rhythm class perception by expert phoneticians. In *Proceedings of 18th International Congress of Phonetic Sciences*. Glasgow.

[16] Repp, B. H. 2003. Rate limits in sensorimotor synchronization with auditory and visual sequences: The synchronization threshold and the benefits and costs of interval subdivision. *J. Mot. Behav. 35,* 355–370.

[17] Repp, B. H. 2005. Sensorimotor Synchronisation; a review of the tapping literature. *Psychon. Bull. Rev. 12,* 969–992.

[18] Repp B.H., Penel A. 2004. Rhythmic movement is attracted more strongly to auditory than to visual rhythms. *Psychol Res. 68(4),* 252-70.

[19] Roach, P. 1982. On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (ed.) *Linguistic controversies*. London, pp. 73-79.

[20] Scott, D. R., Isard, S. D., de Boysson-Bardies, B. 1985. Perceptual isochrony in English and in French. *Journal of Phonetics 13(2),* 155-162.

[21] Šturm, P., Volín, J. 2016. P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment. *Journal of Phonetics 55*, 38–52.

[1] The total duration of [b] was approximately 35 ms in both verbal prompts, i.e. syllable and vowel onsets differed only minimally.

[2] Syllable onset was measured from the onset of voicing in [b].