International Symposium on Performance Science
ISBN 978-94-90306-02-1

Entrainment to speech and song

Pascale Lidji[1,2,3], Caroline Palmer[1], Isabelle Peretz[2], and Michele Morningstar[1]

[1] Department of Psychology, McGill University, Canada
[2] BRAMS Laboratory, University of Montreal, Canada
[3] Department of Psychology, Université Libre de Bruxelles, Belgium

Do we entrain similarly to speech and song? English and French participants were asked to tap along with the same set of sentences, in three conditions that varied in temporal regularity and musicality. The utterances, produced by a single vocalist, were either spoken naturally, spoken regularly (aligning syllables with a metronome), or sung regularly. Participants tapped more regularly to both song and regular speech than to natural speech. One can entrain to natural stimuli that are not musical: participants tapped with similar regularity to regularly spoken and to regularly sung stimuli. However, participants' taps were better aligned with the metronome underlying song than regular speech. Although sensitivity to rhythmic regularities is not unique to music, the current findings support the idea that music, due to its rhythmic structure, is a privileged stimulus to elicit entrainment.

Keywords: entrainment; tapping; speech; song; rhythm

Entrainment is the natural tendency to perceive and to synchronize one's responses with temporal regularities present in external stimuli (Large and Jones 1999). One can entrain to stimuli of various rhythmic complexities, from a steady metronome pulse to complex music exhibiting several metrical levels (Drake *et al.* 2000), to stimuli with no obvious regular beat, such as speech (Wilson and Wilson 2005). Entrainment is generally investigated by using sensorimotor synchronization tasks such as tapping. Tapping variability and synchronization with the underlying beat can be used to infer entrainment strength.

Most documented cases of entrainment are related to temporally regular stimuli (Phillips-Silver *et al.* 2010). Can we entrain to stimuli with no under-

lying steady beat? Research on entrainment to speech suggests that conversational turn-taking (Wilson and Wilson 2005) can be simulated with an oscillator model that has also been applied to entrainment with music (Large and Palmer 2002). The question of whether speech can be considered a regular stimulus is matter of debate because speech is not isochronous (for a review, see Patel 2008). Listeners should entrain better to regular stimuli, such as singing, than to less regular stimuli, such as speech. Another question is whether melodic information is necessary to elicit entrainment. Snyder and Krumhansl (2001) have shown that removing pitch variations in ragtime piano music does not impair listeners' tapping performance; the removal of pitch information has not been examined with human speech or song. Although one could entrain to both speech and song, music (and thus song) might be the paramount stimulus for eliciting synchronized motor responses. Finally, one can wonder whether rhythmic differences between languages influence entrainment. Stress-timed languages, such as English, are usually perceived as more rhythmically regular than syllable-timed languages, such as French (Cutler 1991). Therefore, one could entrain more to English than to French.

We investigated how people entrain to vocal productions that vary in their musicality, temporal regularity, and language rhythmic class. We asked English and French-speaking participants to tap to English and French utterances produced in three conditions: (1) spoken with a natural (irregular) prosody (henceforth, *natural speech*), (2) spoken with a metronome inducing an underlying beat (*regular speech*)—a condition that could be compared to rap music or poetry slam, and (3) *sung* with a metronome. Because the naturally spoken condition has no clear underlying beat, we expected participants to tap more variably with natural speech than with regular speech and song. In addition, preferential entrainment to music would predict that participants should tap less variably and with fewer asynchronies to song than to regular speech.

METHOD

Participants

Twenty-four monolingual native English speakers (mean age=23.1 years, range 19-40) and 24 monolingual native French speakers (mean age=26.5 years, range 18-46) were recruited from the Montreal area. Participants were not selected for their musical experience. The study complied with the norms of the McGill University Ethics Review Board.

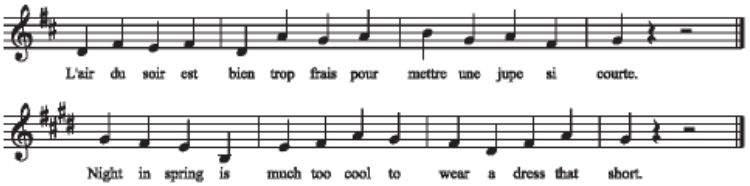

Figure 1. Score and lyrics for one sung stimulus in French with its English counterpart.

Materials

Twelve English and 12 French sentences were recorded in three conditions by a balanced English-French bilingual speaker experienced in singing. All sentences were composed of 13 monosyllabic words. English and French sentences were matched on word frequency, syntactic structure, and rhythmic structure. The three recording conditions were as follows. In the *naturally spoken condition*, the speaker was instructed to speak the sentence with a natural prosody. In the *regularly spoken condition*, the speaker was asked to align every other syllable with a 120 bpm (500 ms interonset interval) metronome click presented through headphones. In the *regularly sung condition*, each sentence was sung *a capella* on a melody of 13 quarter notes (one note per syllable), aligning every other note with a 120 bpm metronome click. Twelve tonal melodies were composed for the sung condition (7 major, 6 minor; see Figure 1). Each sentence was paired with two different melodies in the sung condition. Naturally spoken utterances had an average duration of 3.43 s, regularly spoken utterances 3.83 s, and sung utterances 4.18 s.

Procedure

All auditory stimuli were presented to participants over headphones. Tapping responses were recorded on a silent electronic keyboard as midi data, with a temporal resolution of 1 ms. Participants' spontaneous tapping rate was measured at the beginning of the experiment; they were asked to tap at a regular and comfortable pace with the index finger of their dominant hand, for 30 s. This was followed by the speech, regular speech, and song tapping task, in which participants were instructed to tap along to the beat they perceived in the utterances they heard. On each experimental trial, an utterance (naturally spoken, regularly spoken, or sung) was presented three times. Participants were instructed to listen to the first presentation of the utterance, and to tap along to the stimulus on the second and third repetitions. English- and French-speaking participants were presented with both English and

French stimuli, blocked by language, with the order of language presentation counterbalanced among participants. Within a language block, conditions (natural speech, regular speech, song) were mixed and experimental trials were presented in a pseudo-random order within each language. In the sung condition, the sentence-melody pairing was counterbalanced across participants (each participant heard each sentence paired with only one melody). Between language blocks, participants completed a questionnaire about their linguistic and musical background. At the end of the speech tapping task, the participants completed a second measure of their spontaneous tapping rate. Finally, they were asked to tap along with a sounded metronome (IOI=500 ms, for 30 s) to assess their synchronization accuracy with a simple stimulus. Speech and song tapping data were collected from a total of 144 trials (2 languages x 12 sentences x 3 conditions x 2 repetitions). The participants' language was a between-subjects variable; the stimulus language and the condition were within-subject variables.

RESULTS

There was a main effect of condition on tapping variability in the speech and song tapping task as indexed by the Coefficient of Variation of Inter-Tap Intervals (CV [ITI]; SD/M), $F_{2,92}=45.16$, $p<0.001$. Tukey's post-hoc tests revealed that participants tapped more variably to natural speech (M=0.30, SD=0.15) than to regular speech (M=0.18, SD=0.12) or to song (M=0.12, SD=0.13) (see Figure 2, left panel). Tapping variability was significantly higher for English than for French stimuli, $F_{1,46}=4.21$, $p<0.05$. The same ANOVA run on the Coefficient of Variation of the Inter-Syllabic Intervals (CV [ISI]) of the stimuli similarly revealed a main effect of condition, $F_{2,22}=142.91$, $p<0.001$. Naturally spoken (M=0.46, SD=0.10), regularly spoken (M=0.33, SD=0.08), and sung stimuli (M=0.19, SD=0.05) each differed significantly from each other (see Figure 2, right panel).

Participants' synchronization with the regular stimuli (regular speech and song) was compared by examining the asynchronies of their taps relative to the timing of the nearest metronome click to which the singer had been asked to synchronize her production (not heard by the participants). Participants' taps tended to be anticipatory for both types of stimuli. The asynchronies were smaller to sung stimuli (M=-3.9 ms, SD=33.4) than to regularly spoken stimuli (M=-15.3 ms, SD=36.6), $F_{1,46}=18.1$, $p<0.001$. English and French participants did not differ significantly on any of the control tasks (spontaneous motor tempo and tapping with a metronome), nor did they differ in their performance in the speech and song tapping task.

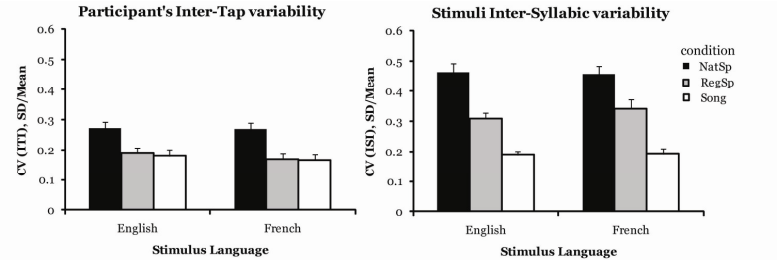

Figure 2. Left panel: Mean Coefficient of Variation (CV) of Inter-Tap Intervals (ITI) by stimulus language and experimental condition (natural speech, regular speech, song). Right panel: Mean Coefficient of Variation (CV) of Inter-Syllabic Intervals (ISI) by stimulus language and experimental condition. Error bars represent the standard error of the mean.

DISCUSSION

English and French participants were asked to tap with vocal utterances that exhibited a gradient of temporal regularity from natural speech (less regular) to song (most regular). The presence of an underlying beat facilitated entrainment in regular speech and song compared with vocal stimuli with no regular pulse (natural speech). Listeners tapped with a similar regularity to regularly spoken and to sung utterances, even though the syllables in the stimuli were spaced more regularly in song than in regular speech. However, tapping asynchronies were smaller for song than for regular speech. Our results suggest that temporal regularity can be extracted from vocal (verbal) utterances that convey a rhythmic pulse but have no melodic variations. Our regularly spoken stimuli were comparable to poetry or rap-music, suggesting that one can entrain to speech if it is regularized in a musical way. This supports the idea that music, due to its rhythmic structure, is a privileged stimulus to elicit entrainment. The present findings generalize across English and French speakers, but participants tapped more variably to English than to French utterances. This seems to contradict the idea that it is easier to entrain to stress-timed than syllable-timed languages; at least when spoken utterances are mixed with sung utterances. Further research should generalize our conclusions to a larger sample of speakers, as well as to less controlled situations and stimuli, such as people tapping or clapping their hands to real-world singing and rap music.

Acknowledgments

This research was supported by postdoctoral fellowships from the Belgian FRS-FNRS and the WBI-World Program to the first author, and a Canada Research Chair and NSERC grant 298173 to the second author.

Address for correspondence

Pascale Lidji, Department of Psychology, McGill University, 1205 Doctor Penfield Avenue, Montreal, Quebec H3A 1B1, Canada; *Email:* pascale.lidji@mcgill.ca

References

Cutler A. (1991). Linguistic rhythm and speech segmentation. In J. Sundberg, L. Nord, and R. Carlson (eds.), *Music, Language, Speech, and Brain* (pp. 157-166). London: Macmillan.

Drake C., Penel A., and Bigand E. (2000). Tapping in time with mechanically and expressively performed music. *Music Perception*, *18,* pp. 1-23.

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

Large E. and Jones M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106,* pp. 119-159.

Patel A. D. (2008). *Music, Language, and the Brain*. Oxford: Oxford University Press.

Phillips-Silver J., Aktipis C. A., and Bryant G. A. (2010). The ecology of entrainment: Foundations of coordinated rhythmic movements. *Music Perception, 28*, pp. 3-14.

Snyder J. and Krumhansl C. C. (2001). Tapping to ragtime: Cues to pulse finding. *Music Perception*, *18,* pp. 455-489.

Wilson M. and Wilson T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review, 12*, pp. 957-968.