EFFECTS OF RHYTHMICITY ON SPEECH PERCEPTION IN SPEECH AND MUSICAL CONTEXTS

Kathryn Franich & Sreeparna Sarkar

University of Delaware
kfranich@udel.edu, sree@udel.edu

ABSTRACT

Speech and music share the property of being organized into hierarchical units which influence aspects of their timing and perception. A key difference is found in the specific timing patterns found in language and music, however, in that the latter generally shows a much stricter adherence to rhythmic regularity. The higher level of precision in processing of temporal and melodic events in music has been hypothesized to be a key factor in explaining how musical experience can lead to enhanced speech processing [24]. Consistent with this hypothesis, we show that perception of speech in the context of musical beats is more adversely affected by contextual rhythmic irregularities than is speech in the context of other nonsense speech sounds or within a cohesive phrase. In some phrasal contexts, temporal irregularities are in fact found to be beneficial for perception.

Keywords: Rhythm, timing, speech perception, metrical structure, music

1. INTRODUCTION

Speech and music share many phonetic and structural attributes, including use of pitch and timing patterns to signal grouping of hierarchically-organized phrasal units [3]. From a perceptual standpoint, performance in both domains is also demonstrated to rely on the performer/listener's ability to use preceding timing patterns to predict upcoming events [14,17,27,29, 30]. For example, it is shown that listeners/performers are attuned to temporal regularities in the signal and use these regularities to predict the location of upcoming beats or syllables [14,27]. Timing patterns across the two domains are often quite different, however: while spoken language may in some instances display evidence of relative periodicity in the timing of syllables or stress feet [8,12], overall, crosslinguistic patterns of speech timing indicate relatively little evidence for periodicity in speech [20], whereas it is quite common to find such periodicities in many types of music [7]. It has been argued that the relatively higher level of melodic and temporal precision found in music is attributable to its function as a medium for joint action and as source of emotional reward, quite different from the communicative function of speech [3,24,25]. It is thus predicted that there is a higher demand on attentional resources during music perception as compared with speech perception, a hypothesis which is supported by fMRI research showing enhanced activation in the same brain regions in response to speech perceived as song versus that perceived simply as speech [33]. Furthermore, a body of recent work shows that the different level of attentional detail required in music may have consequences for speech processing, as individuals with musical training have been found to have enhanced speech perception abilities [5,21,22], and music-based intervention methods have been shown to positively impact speech development in individuals with a variety of speech and language disorders [10,34].

So far, temporal processing across the domains of speech and music has received relatively less attention than studies on pitch and melody. This is in spite of the fact that much recent work has focused on the role of neural entrainment to temporal regularities in the speech signal and its role in language perception, acquisition, and rehabilitation [10,11,26]. The present study aims to fill this gap by exploring how temporal expectations may vary across more and less musical contexts. Specifically, we ask whether listeners, regardless of music experience, show more fine-grained temporal predictions for speech when presented in a more musical context (following a drumbeat) versus in a speech context.

2. METHOD

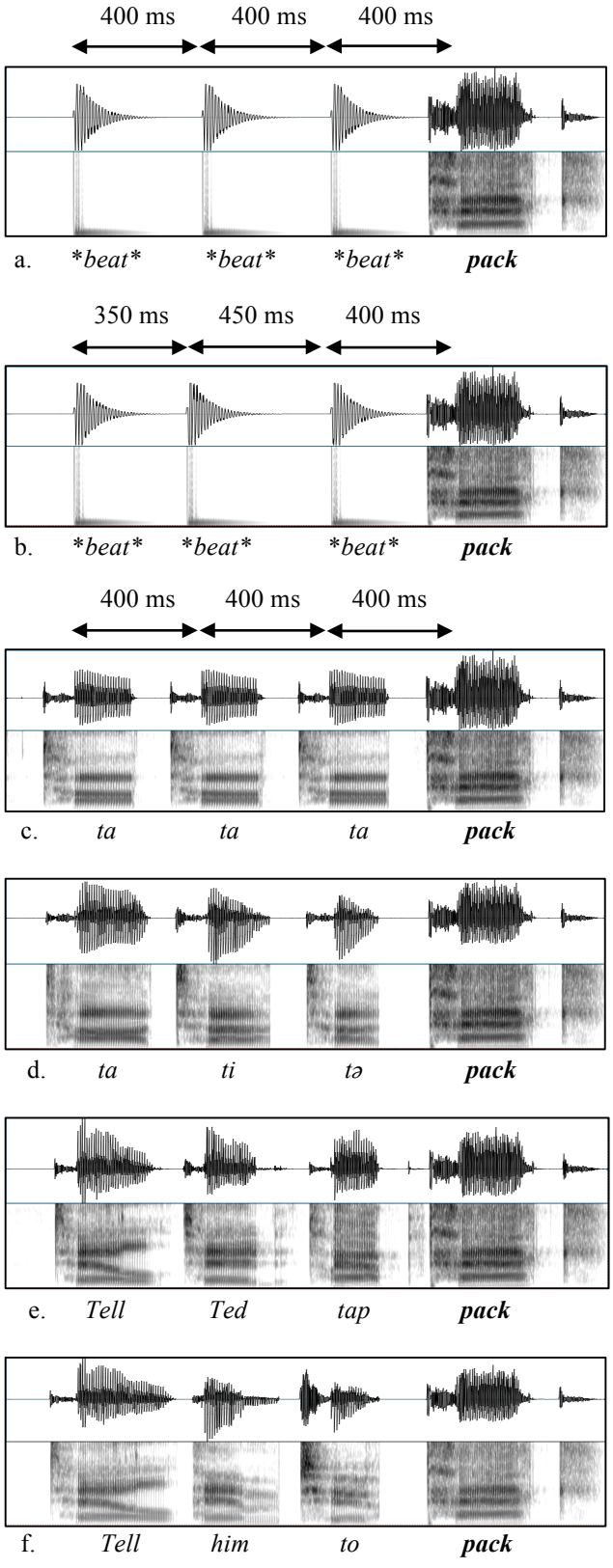
The experiment consisted of various sets of stimuli in which three context beats or syllables were played before one of two target words. The goal of the task was to identify the target word as soon as possible. In the DRUMBEAT condition (Fig. 1a,b), the three context beats were created using the Risset Drum synthetic beat in Audacity [1] software, v2.1.2. The drum beats were modified to have a 300 second duration/decay and a center frequency of 120 Hz. Width of the noise band was set to 1000 Hz. Three of these beats were presented such that the duration between beat onsets was 400 milliseconds. The target word (either *pack* or *mop*), recorded by a male native speaker of English, was then positioned such that the

onset of its vowel (a close approximation to the word's 'p-center' [18,31]) was another 400 ms from the onset of the final drumbeat. In the SYLLABLE condition (Fig. 1c), a nonsense syllable *ta* was recorded by the same speaker whose total duration was 300 ms (with a vowel duration of 200 ms). This syllable was chosen due to its sound-symbolic nature, as it is frequently used in western musical tradition for mimicking a drumbeat or metronome (as in, for example, the Kodály method [6]). As with all speech conditions, F0 of the vowel of the context syllables was flattened to 120 Hz, the speaker's average F0 for the recorded utterances, using the Pitch-Synchronous Overlap-and-Add (PSOLA) algorithm in Praat [4].[1] Additionally, for all speech conditions, consecutive context syllables and the target word were positioned such that their respective vowel onsets/p-centers were 400 ms apart. In the SENTENCE condition (Fig. 1e), the three context syllables constituted a cohesive phrase, *Tell Ted tap…* The phrase was spoken naturally by the speaker and therefore displayed some variability in syllable duration, though all three words were close to 330 ms in duration. Sonorous portions of the rhyme for the three words were 300 ms, 170 ms, and 150 ms, respectively. In the SENTENCE_PWORD_INTERNAL condition (Fig. 1f), stimuli again formed a sentence, this time with the second two syllables replaced with two function words for *Tell him to…*, which together form a single prosodic word [32] or clitic group [19]. Again, syllables varied in duration, with an average duration of around 310 ms and sonorous rhyme durations of 300 ms, 240 ms, and 150 ms. Finally, two additional conditions were incorporated in order to mimic the durational and spectral variability of the two sentence conditions within the context of nonsense syllables. The SYLL_DURATION_VARIED involved the same syllable *ta* as in the SYLLABLE condition, but with its duration manipulated to vary from 300 ms, to 200 ms, to 150 ms. Finally, the SYLL_VOWEL_VARIED condition (Fig. 1d) consisted of of the syllables *ta ti tə*, again with durations manipulated to 300 ms, 200 ms, and 150 ms on successive syllables. Stimulus amplitudes were normalized to 65 Hz.

2.1 Isochrony Manipulation

For all stimulus conditions, four additional manipulations were carried out such that the resulting stimulus strings were made increasingly (though subtly) less *isochronous* (or temporally regular) by removing either 25 or 50 ms of silence from between the first two beats/syllables and adding it to the silent portion between the second and third beats/syllables (the SHORT-LONG condition) or the reverse: removing

Figure 1: Annotated acoustic signals for the isochronous drumbeat (a), syllable (c), syllable-vowel varied (d), sentence (e), & sentence-pword internal (f) conditions, and drumbeat condition with 100 ms deviation, short-long (b). Arrows indicate duration between p-centers used for isochrony/ deviation calculation in drum and speech conditions.



silence from between the second two beats and adding it to the silence between the first two beats (the LONG-SHORT condition). Thus, the resulting sequences had either a 50 ms or a 100 ms difference in duration (or 'deviation') between the first pair and the second pair of syllables. These manipulations resulted in five different possible rhythmic variations: 0 ms change; short-long with 50 ms difference; short-long with 100 ms difference; long-short with 50 ms difference; long-short with 100 ms difference. Two untrained listeners rated the forms with the 100 ms deviations as sounding the most arrhythmic, and those with 50 ms deviations as sounding slightly less rhythmic than the unaltered forms.

All of these manipulations resulted in a 6 (context) x 5 (rhythm manipulation) x 2 (target word) design, for a total of 60 distinct stimuli. The experiment was blocked by context, with blocks presented in random order; each block included 6 repetitions of each stimulus for a total of 360 stimuli heard during the course of the experiment.

2.2 Participants and Procedure

50 participants (40 recruited from Amazon Mechanical Turk and 10 recruited in the University of Delaware Phonetics Lab) aged 19-71 (mean age = 32) participated in a forced choice task in which they were asked to identify, as quickly as possible, the last word of the sentence they had heard by pressing the 0 or 1 keys on the keyboard. The experiment was web-based, developed using jsPsych [9] and administered through JATOS [16]. Participants were given a short break between each block of the experiment. After the experiment, they completed a brief demographic survey which included questions about their language background and music experience. 17 out of the 50 participants had some level of musical training, which ranged from a year or less to ongoing participation in music lessons or ensemble play. No participant reported any hearing impairment.

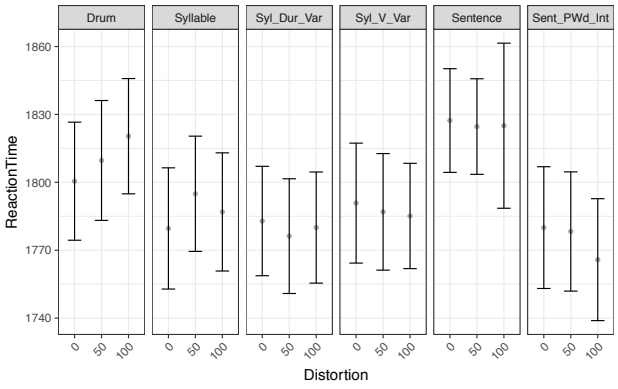
2.3 Hypotheses

It is hypothesized that deviations from isochrony will have a more profound negative effect on perception within the DRUMBEAT context, such that smaller (50 ms) and larger (100 ms) deviations will lead to increased reaction times on identification of the target word. It is furthermore predicted that the three SYLLABLE conditions (plain, duration varied, and vowel varied) will show an effect of the larger (100 ms) deviation from isochrony, but perhaps not the smaller (50 ms) deviation. Finally, it is predicted that the two SENTENCE conditions will show the least impact of the rhythmic deviations.

3. RESULTS

Three participants' data were removed from the dataset as their reaction times were more than 2 standard deviations above the mean, indicating they were likely not adhering to the instruction to respond as quickly as they could. For the remaining subjects, data was modelled using linear mixed effects models with the *lmer* package [2] for *R* statistical software [28]. The full model included categorical fixed effects of CONTEXT (6 levels; see Fig. 1), DEVIATION (3 levels: 0, 50 ms, 100 ms), DIRECTION (2 levels: long-short or short-long), and MUSIC EXPERIENCE (2 levels: yes or no), as well as TRIAL ORDER, coded as a continuous variable and mean-centered. Factors CONTEXT and DEVIATION were treatment-coded, with reference levels set to the drumbeat condition and the 0 ms deviation condition, respectively; the other two factors were sum-coded. By-subject random slopes were included for each factor. Significance was determined based on Satterthwaite's degrees of freedom with the lmerTest package [15] for *R*.

Figure 2: Reaction times across stimulus context and rhythm deviations for the short-long condition



Results revealed no effect of any level of deviation from isochrony in the long-short direction ($p > .25$). To simplify the analysis, the data was subset to only compare across deviation levels in the short-long direction. Results revealed a main effect of CONTEXT on reaction time, with slightly longer reaction times recorded for the SENTENCE context (*Tell Ted tap…*) than the other contexts (β=40.57; t=3.71; $p<0.001$). There was also a main effect of DEVIATION for the 100 ms condition, such that participants were overall slower to respond in the condition with the rhythmic deviation than in the perfectly rhythmic (0 ms deviation) condition (β=15.91; t=2.03; $p<0.05$). There was no main effect of deviation for the 50 ms condition ($p = 0.41$). As can be seen in Figure 2, there was also a significant interaction between CONTEXT and DEVIATION for the 100 ms condition: the SENTENCE_PWORD_INTERNAL (*Tell him to…*) context showed the opposite effect of rhythmic

deviation, as participants were faster to respond when the 100 ms deviation was present (β=-24.55; t=-2.26; p<0.05). This interaction between CONTEXT and DEVIATION trended in the same direction (but did not reach significance) for the SENTENCE context (*Tell Ted tap…*) (β =-20.65, t =-1.65, p=.24), and two of the syllable contexts: the SYLL_DURATION_VARIED context (*ta ta ta…*, duration variable) (β=-20.265, t=-1.909, p=.056) and the SYLL_VOWEL_VARIED context (*ta ti tə…*, duration variable) (β=-20.235, t=-1.92, p=0.055). This interaction was not significant for the plain SYLLABLE context (*ta ta ta…*, duration equal) (p>.30), where, similar to the DRUMBEAT condition, reaction times were numerically slower in both of the two deviation conditions as opposed to the perfectly rhythmic condition. To get a better picture of how the deviation conditions were affecting perception in the DRUMBEAT and plain SYLLABLE contexts, which appear to pattern most similarly to one another, individual models were constructed for the two context conditions which included fixed effects of DEVIATION, MUSIC EXPERIENCE, and TRIAL ORDER and by-subject random slopes for all factors. For the DRUMBEAT condition, results revealed significantly longer reaction times in the 100 ms deviation condition when compared to the perfectly rhythmic condition (β=8.343, t=2.28, p<0.05). Reaction times were intermediate in the 50 ms deviation condition, but not significantly longer than they were in the perfectly rhythmic condition (β=3.04, t=.837, p=0.40). For the SYLLABLE context, the difference in reaction times between the 100 ms and perfectly rhythmic conditions did not reach significance (β=2.657, t=1.28, p=.20), nor did it reach significance between the 50 ms and perfectly rhythmic conditions (β=6.917, t=1.56, p=.13). There was no effect of MUSIC EXPERIENCE for either dataset (p > .30).

4. DISCUSSION

The results from the present study show that listeners' temporal predictions about an upcoming word are more sensitive to deviations from temporal regularity in a context sequence containing drumbeats, as opposed to a sequence of nonword speech syllables, or within a coherent phrase. These results are consistent with accounts positing a greater allocation of attentional resources in perception during music than during speech. Since the non-speech drumbeat was meant to encourage participants to enter a more musical mode of perception, we expected temporal expectations in this condition to be more fine-grained. Also of interest was the fact that temporal deviations within a sentence actually *facilitated* target word perception—this effect could be seen most clearly in the case of the sentential context *Tell him to…* where the syllables preceding the target word form a single prosodic unit. This finding is not surprising, given the reduced duration between the initial two syllables *Tell him* in the deviation conditions allows them to more closely follow the pattern of temporal compression commonly found within binary feet in English [13]. Note that this effect cannot be attributed solely to the greater level of durational and spectral variability among the context syllables in this condition, as a comparable facilitative effect of the 100 ms deviation was not found for the SYLL_DURATION_VARIED & SYLL_VOWEL_VARIED conditions. It is interesting to note, however, that these two context conditions did seem to pattern rather differently from the evenly timed, non-variable SYLLABLE condition. For example, reaction times were numerically shorter in the perfectly rhythmic condition for the SYLLABLE context, as was found for the DRUMBEAT context, whereas the SYLL_DURATION_VARIED & SYLL_VOWEL_VARIED contexts showed numerically slower reaction times in the rhythmically-even condition as opposed to the deviation conditions. This suggests that the variability of vowel durations in the latter two contexts did somewhat affect listener predictions about the timing of the target word.

A possible limitation of the present study is the acoustic difference between the context sounds heard in the drumbeat condition and those heard in the other conditions. Though efforts were made to control for factors such as frequency/F0 and stimulus duration, there are still elements which clearly distinguish the drumbeat stimuli from the speech stimuli, such as the faster rate of decay of the sound over time for the drumbeat. Thus, it could be that sensitivity to rhythmic deviations arose simply because of the nature of the stimulus, and not because the drumbeat is a more musical stimulus, per se.

Either way, our results clearly show that predictions about timing are influenced by language structure (e.g. prosodic word-internal vs. external patterns), and not driven purely by rhythmic regularity in the stimulus. Mapping out how, exactly, these kinds of linguistic knowledge interact with other kinds of lower-level expectations about temporal regularities will be key to understanding of how temporal predictions for language are constructed. Furthermore, the musical context used here was extremely simple, whereas music often displays variation in timing which may be more comparable to speech. Future work should explore how predictions in the context of highly rhythmic music relate to those made in more complex, fluid rhythmic musical contexts.

5. REFERENCES

[1] Audacity Team 2018. Audacity(R): Free Audio Editor and Recorder [Computer application]. V.2.1.2 ret. Jan 2016 from https://audacityteam.org/.

[2] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Soft.* 67,1, 1-48

[3] Besson, M., Chobet, J., Marie, C. 2011. Transfer of training between music and speech: Common processing, attention, & memory. *Front. Psych.* 2, 94.

[4] Boersma, P., Weenink, D. 2017. Praat: doing phonetics by computer [Computer program]. Version 6.0.29, retrieved 2017 from http://www.praat.org/

[5] Chobert J., François C., Velay J. L., Besson M. 2014. Twelve months of active musical training in 8 to 10 year-old children enhances the preattentive processing of syllabic duration and voice onset time. *Cereb. Cortex* 24, 4, 956-967.

[6] Choksy, L. 1999. *The Kodály method I: comprehensive music education.* 3rd ed. Upper Saddle River, N.J.: Prentice Hall.

[7] Cooper, G., Meyer, L.B. 1960. *The rhythmic structure of music*. Chicago: U. Chicago Press.

[8] Cummins, F., R. Port. 1998. Rhythmic constraints on stress timing in English. *JPhon* 26, 145-71.

[9] de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Meth.* 47(1), 1-12.

[10] Fijii, S., Wan, C.Y. 2014. The role of rhythm in speech and language rehabilitation: The SEP hypothesis. *Front. Hum. Neuro.* 8, 777.

[11] Giraud, A., Poeppel, D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neuro.* 15, 4, 511-517.

[12] Hawkins S. 2014. Situational influences on rhythmicity in speech, music, and their interaction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369,1658, 20130398.

[13] Huggins, A.W. F. 1972. On the perception of temporal phenomena in speech, *JASA* 51, p. 1279-1290.

[14] Jones, M.R., Moynihan, H., MacKenzie, N., Puente, J. 2002. Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychol. Sci.* 13, 4, 313-319.

[15] Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B. 2017. "lmerTest Package: Tests in Linear Mixed Effects Models." *J. Stat. Soft.* 82,13, pp. 1-26.

[16] Lange K, Kühn S, Filevich E (2015) "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLoS ONE* 10,6, e0130834.

[17] Large, E., Jones, M.R. 1999. The dynamics of attending: How people track time-varying events. *Psychol. Rev.* 106,1, 119–159.

[18] Morton, J., Marcus, S., Frankish, C. 1976. Perceptual centers (p-centers). *Psychol. Rev. 83*. 405–408.

[19] Nespor, M., Vogel, I. 1986. *Prosodic Phonology*. Dordrecht: Foris Publications.

[20] Nolan, F. and Jeon, H-S. 2014. Speech rhythm: a metaphor? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 19, 369, 1658.

[21] Ott, C. G. M., Langer, N., Oechslin, M., Meyer, M., Jäncke, L. 2011. Processing of voiced and unvoiced acoustic stimuli in musicians. *Front. Psychol*. 2, 195.

[22] Parbery-Clark, A., Strait, D., Kraus, N. 2011. Context-dependent encoding in the auditory brainstem subserves enhanced speech-in-noise perception in musicians. *Neuropsychologia* 49, 12, 3338-3345.

[23] Patel, A.D. 2003. *Music, Language, and the Brain*. Oxford: Oxford University Press.

[24] Patel, A.D. 2011. Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Front. Psych.* 2, 142.

[25] Patel, A.D., 2012. The OPERA hypothesis: assumptions and clarifications. *Ann. N. Y. Acad. Sci*. 1252, 124e128.

[26] Power AJ, Mead N, Barnes L, Goswami U 2012 Neural entrainment to rhythmically presented auditory, visual, and audio-visual speech in children. *Front. Psych.* 3, 216.

[27] Quené, H., Port, R. F. 2005. Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica* 62, 1-13.

[28] R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[29] Rajendran, V.G., Teki, S., Schnupp, J.W.H. 2018. Temporal processing in audition: Insights from music. *Neuroscience* 1, 389, 4-18.

[30] Rankin, S., Fink, P.W., Large, E.W. 2014. Fractal structure enables temporal prediction in music. *JASA* 136, EL256.

[31] Scott, S. 1998. The point of P-centres. *Psychological Research* 61, 4–11.

[32] Selkirk, E.O. 1995. The prosodic structure of function words. *UMOP 18*. GLSA, UMass, Amherst.

[33] Tierney A., Dick F., Deutsch D., Sereno M. 2013. Speech versus song: multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cereb. Cortex* 23, 249–254

[34] Wan C. Y., Demaine K., Zipse L., Norton A., Schlaug G. 2010. From music making to speaking: engaging the mirror neuron system in autism. *Brain Res. Bull.* 82, 161–168.

[1] Given the speaker already had a relatively monotone voice (F0 for the unaltered utterances didn't vary more than 5 Hz above or below the mean), PSOLA manipulations did not result in abnormally monotone-sounding stimuli.