Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, July 23-28, 2012, Thessaloniki, Greece
Cambouropoulos E., Tsougras C., Mavromatis P., Pastiadis K. (Editors)

How Fast Can Music and Speech Be Perceived? Key Identification in Time-Compressed Music with Periodic Insertions of Silence

Morwaread M. Farbood,[*1] Oded Ghitza,[#2] Jess Rowland, [§3] Gary Marcus, [‡4] David Poeppel, [‡†5]

[*] *Dept. of Music and Performing Arts Professions, New York University, USA*
[#] *Dept. of Biomedical Engineering, Boston University, USA*
[‡] *Dept. of Psychology, New York University, USA*
[§] *Dept. of Art Practice, University of California, Berkeley, USA*
[†] *Center for Neural Science, New York University, USA*

[1]mfarbood@nyu.edu, [2]jessrowland@berkeley.edu, [3]oghitza@bu.edu, [4]gary.marcus@nyu.edu, [5]david.poeppel@nyu.edu

ABSTRACT

Background

This study builds upon recent work that indicates the overlap of neural resources in the processing of music and speech (Federenko, Patel, Casasanto, Winawer, & Gibson, 2009; Koelsch, Gunter, Wittfoth, & Sammler, 2005; Kraus & Chandrasekaran, 2010; Patel, 2008). It is based on an experimental design used by Ghitza and Greenberg (2009) in an experiment that explored the possible role of brain rhythms in speech perception by inserting silences in compressed speech and ascertaining the error rate at identifying words. They used semantically unpredictable (nonsensical) but grammatical sentences that were compressed to three times the speed of normal speech. Without silences added, the error rate for word identification was >50%. However, when silences (up to 160 ms) were added between every 40 ms segment of audio, performance improved, resulting in a U-shaped error rate curve with a preferred *packaging rate* of around 6 to 17 Hz. Packaging rate is defined as the periodic silence-plus-audio-segment rate.

Ghitza and Greenberg interpreted the decrease in error rate corresponding to the insertions of silence as essentially the result of adding "necessary" decoding time. Based on these results, they implicated an oscillatory mechanism on a specific timescale for auditory processing.

Aims

This work addresses whether the perception of musical structure is subject to a similar principle as speech. Potential parallels would indicate possible shared mechanisms between these two domains, and the study of these oscillatory mechanisms may further open up new avenues of research into basic psychoacoustic processing.

The strategy employed in this study was to apply Ghitza and Greenberg's gap-insertion experimental paradigm to time-compressed melodic sequences. Instead of identifying words, the task was to identify the *key* of a melody. The goal was to discover whether the U-shaped error rate curve found for speech had an analog in music, and if such a curve were present, what precisely was the decoding time needed for music processing.

Method

The 10 melodic sequences for this experiment were based on stimuli composed for a previous study by Farbood, Marcus, Mavromatis, and Poeppel (2010) that explored the psychophysics of structural key-finding. In this study, musically trained subjects were asked to judge whether melodic sequences presented at different tempos ended on a resolved or unresolved pitch. In the study described here, the sequences were first time-compressed then altered by inserting varying durations of silences between audio segments. The sequences were compressed to 1680 bpm (28 Hz), fast enough to make key identification impossible. The waveforms for the compressed sequences were segmented into consecutive audio chunks of equal duration, each followed by a silence gap. The independent parameters were the duration of the audio segments (10 to 65 ms) and the duration of the silence gap (40 to 1280 ms).

The participants consisted of 28 musically trained listeners (average age 23.64 years, *SD* = 5.73, 25 male). Formal training on a primary instrument was an average of 9.63 years (*SD* = 4.84). On a scale of 0 to 5 (where 0 was no musical experience and 5 was professional-level musical experience, subjects' mean self-ranking was 3.77 (*SD* = 0.75). Average number of years of college-level music theory was 2.07 (*SD* = 1.65), and two subjects reported having absolute pitch.

Participants asked to indicate whether each sequence sounded resolved (ending on an implied tonic) or unresolved (ending on an implied dominant) by entering responses into a computer interface. Subjects were instructed to ignore aspects such as perceived rhythmic or metrical stability when making their decision.

Each participant listened to 340 sequences: each of the 10 sequences *twice* at an uncompressed tempo of 60 bpm/1 Hz without silences inserted and compressed rate of 1680 bpm/28 Hz also without silences inserted, plus the 10 sequences altered at all combinations of the audio segment and silence durations *once* (5 audio segment durations x 6 silence durations x 10 sequences). Stimuli were presented in a pseudorandomized order that took into account tempo, key, and original sequence, such that no stimulus was preceded by another stimulus generated from the same original sequence or of the same type (uncompressed, compressed without gaps, and compressed with gaps), and no stimulus was in the same key as the two preceding sequences. All stimuli were

transposed such that they were at least three sharps/flats away from the key of the immediately preceding stimulus. The experiment took approximately one hour to complete.

Two strategies were used to determine response accuracy: the first method entailed labeling a response as correct if it matched the empirical key judgments from Farbood et al. (2010), and the second method entailed determining correctness by looking at the each subject's judgments on the original, unmodified sequences played at the optimal tempo (60 bpm). Both strategies resulted in the same findings.

Results

Absent silences, listeners were at chance in discerning musical key in the high tempo (1680bpm) sequences. Once silences were inserted, however, subjects achieved above chance performance. A two-factor, repeated-measures ANOVA and post-hoc tests revealed a strong interaction between audio segment and silence durations: there was little improvement at the shortest audio segment sizes (10 and 23 ms), regardless of the length of the inserted silence, but marked improvement for longer audio segments when silences of 160 ms or greater were inserted. For audio segments of 38 ms or longer, a U-shaped error curve was found across silence durations.

The results indicated that insertions of 160-640 ms of silence between audio segments of compressed music significantly reduced error rate. This comes with the caveat that the segments must be long enough in duration to enable pitch to be clearly discerned. Overall, this translates to a preferred *packaging rate*—periodic cycles of audio and silence as described in Ghitza and Greenberg (2009)—for music of 1.4 to 5.1 Hz (65 ms/640 ms audio/silence at the low end, 38 ms/160 ms audio/silence at the high end). This preferred packaging rate is in general agreement with the preferred 0.5 to 7 Hz note event rate (in this case there are eight notes per sequence) estimated by Farbood et al. (2010). However, it suggests significantly shorter decoding time for speech (6-17 Hz) than music. The observed *duty cycle*—audio to silence ratio—for music processing from these results are 24% (for 5.1 Hz) to 10% (for 1.4 Hz) as opposed to 66% (for 17 Hz) to 33% (for 6 Hz) found in speech. In summary, these results thus suggest faster oscillations with a larger duty cycle for speech than music.

Nonetheless, similarities are evident with regard similar error rate curve shapes that were observed both here and in Ghitza and Greenberg (2009). The case for music appears to be slightly more-open ended for the longer silence durations; however, even the longest silence insertions (1280 ms) result in a statistically significant improvement from no silence insertion. Although the U-shape is evident, it is not fully "closed" for lower rates. This is supported by the results of Farbood et. al (2010) in which key judgment accuracy dropped off only below 0.5 Hz (30 bpm). For 38-65 ms audio segment durations, a packaging rate of 0.5 Hz would entail inserting silences that are approximately 1900 ms, which goes considerably beyond the longest duration explored here (1280 ms). The similarities and differences between the results shown here and Ghitza and Greenberg's (2009) results point to the manner in which auditory processing is tied to the specific temporal structure of the input.

Conclusions

Based on these results, emerging questions present themselves: How and why are the duty cycles different for speech and music? How is auditory processing tied to the specific temporal structure of the input? Can we further specify the temporal dynamics of an oscillatory mechanism for music as well as speech processing?

Keywords

Decoding time, neuronal oscillations, speech versus music, key-finding

REFERENCES

Farbood, M., Marcus, G., Mavromatis, P., & Poeppel, D. (2010). The effect of structure and rate variation on key-finding. In *Proceedings of the 11th International Conference on Music Perception and Cognition*. Seattle, USA.

Federenko, E., Patel, A., Casasanto, D., Winawer, J., and Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition, 37*, 1-9.

Ghitza, O. & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception. Phonetica, 66, 113-126.

Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: An ERP study. *Journal of Cognitive Neuroscience, 17*, 1565-1577.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience, 11*, 599–605.

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