Farbood, M. M., Rowland, J., Marcus, G., Ghitza, O., & Poeppel, D. (2015). Decoding time for the identification of musical key. *Attention, Perception & Psychophysics, 77*(1), 28-35.

Decoding Time for the Identification of Musical Key

Morwaread M. Farbood[1], Jess Rowland[1], Gary Marcus[1], Oded Ghitza[2], David Poeppel[1 3]

[1]New York University
[2]Boston University
[3]Max-Planck-Institute (MPIEA), Frankfurt, Germany

Author Note

Morwaread M. Farbood, Dept. of Music and Performing Arts Professions, New York University; Oded Ghitza, Dept. of Biomedical Engineering & Hearing Research Center, Boston University; Jess Rowland, Dept. of Psychology, New York University; Gary Marcus, Dept. of Psychology, New York University; David Poeppel, Dept. of Psychology and Center for Neural Science, New York University and Max-Planck-Institute, Frankfurt, Germany.

This work is supported by NIH 2R01 05660 awarded to DP. Correspondence should be addressed to Morwaread Farbood, Department of Music and Performing Arts Professions, 35 W. 4th St., Suite 1077, New York, NY 10012. E-mail: mfarbood@nyu.edu

Abstract

This study examines the decoding time at which the brain processes structural information in music and compares them to timescales implicated in recent work on speech. Combining an experimental paradigm based on Ghitza and Greenberg (2009) for speech with the approach of Farbood et al. (2013) for musical key-finding, listeners were asked to judge the key of short melodic sequences that were presented at a highly a compressed rate with varying durations of silence inserted in a periodic manner in the audio signal. The distorted audio signals comprised of signal-silence alternations show error rate curves that identify peak performance centered around an event rate of 5–7 Hz (143–200 ms interonset interval; 300–420 beats per minute), where event rate is defined as the average rate of pitch change. The data support the hypothesis that the perceptual analysis of music entails the processes of parsing the signal into chunks of the appropriate temporal granularity and decoding the signal for recognition. The music-speech comparison points to similarities in how auditory processing builds on the specific temporal structure of the input, and how that structure interacts with the internal temporal dynamics of the neural mechanisms underpinning perception.

Keywords: key finding, tonal induction, neuronal oscillations, music structure, brain rhythms, speech rate

Decoding Time for the Identification of Musical Key

Traditionally, most approaches to the perceptual analysis of speech have focused on the rich frequency structure of the signal within a short time window. Speech perception has been—appropriately—characterized as a demanding spectral analysis challenge, and considerable progress has been made investigating the mechanisms underlying short-term frequency analysis (Gold, Morgan, & Ellis, 2011; Stevens, 1998, 2005). Prior work has examined how the temporal structure of speech signals underpins perception in concert with the spectral information (see Rosen, 1992 for review; Drullman, Festen, & Plomp, 1994; Houtgast & Steeneken, 1985; Shannon, Zeng, Wygonski, Kamath, & Ekelid, 1995). One of the emerging generalizations from this line of research is that there appears to be a fortuitous alignment between robust temporal properties of speech, e.g., the envelope fluctuations characteristic of the flow of syllabic information, and the brain rhythms argued to play a role in perception and cognition (Ghitza, 2011; Giraud & Poeppel, 2012; Poeppel 2003). Although the precise mechanisms remain under vigorous debate, there is consensus that both structure in time and processing rate itself merit deeper investigation.

In the theoretical and experimental study of music, there is a long and productive tradition of studying temporal structure and tempo (see London, 2012 for review). However, those approaches have not intersected in principled ways with related speech perception research. Here we capitalize on recent progress in both domains, combining novel approaches to temporal constraints on speech decoding (Ghitza & Greenberg, 2009; Ghitza, 2011, 2012) with results on music perception, and in particular the analysis of key (Farbood, Marcus, & Poeppel, 2013).

The current study builds on an experimental design by Ghitza and Greenberg (2009) that explored the possible role of brain rhythms in speech perception. They inserted periodically spaced silences into semantically unpredictable sentences that were compressed by a factor of three, and measured the error rate in word identification. Without inserted silent gaps, the error rate for word identification in compressed speech was >50%. However, when silence intervals of varying durations (up to 160 ms) were added in between 40 ms segments of audio signal, performance improved, resulting in a U-shaped error-rate curve with a preferred *packaging rate* of around 6–17 Hz (59–167 ms IOI). Packaging rate is a term Ghitza (2011) uses to describe the periodic silence-plus-audio-segment rate of compressed stimuli distorted by silence insertions. For example, stimuli with audio segments of 40 ms and silence intervals of 80 ms would have a 120 ms packaging rate (8.33 Hz). Ghitza and Greenberg (2009) interpreted the decrease in error rate resulting from the insertions of silence as the result of adding necessary decoding time. Based on these results, they suggested an oscillatory mechanism on a specific timescale for auditory processing and developed a phenomenological model to account for these counterintuitive data (Ghitza, 2011).

The association between temporal properties of speech (e.g., mean syllable duration, phoneme duration, etc.) and neuronal oscillations was made explicit by Poeppel (2003) and has subsequently been investigated empirically and computationally in a number of psychophysical and neurophysiological studies (for review, see Giraud & Poeppel, 2012). An important computational angle was introduced by Ghitza (2011, 2013) in the context of formulating a model designed to address how speech signals are

parsed into coarser, typically syllable-long speech fragments, and then decoded. It has now been demonstrated convincingly (Ghitza, 2012) that lower-frequency, theta oscillations are implicated in connected speech parsing; current research is addressing the role of higher frequency beta and gamma oscillations for decoding. Musical stimuli such those used as in the current study have not been explored in this theoretical context, but such materials can help shed light on the mechanistic role that neuronal oscillations might play in perception.

In a study exploring the psychophysics of structural key-finding by Farbood, et al. (2013), the influence of rate variation in music was examined by asking musically trained listeners to judge whether melodic sequences presented at different tempi ended on a resolved or unresolved pitch. The tempi of the sequences were parametrically varied over note event rates of 0.12–56.7 Hz/18–8333 ms interonset interval (IOI)/7–3400 beats per minute (BPM), in which the duration of each note was considered a beat. Error rates on the task resulted in a U-shaped curve where the lowest rates ranged between 30–400 BPM (0.5–6.7 Hz/150–2000 ms IOI). The upper end of the curve overlapped with the range for optimal speech intelligibility and almost precisely aligned with the range in which beat induction and melody recognition occur.

However, a critical unresolved question remained: although it appeared from the results of Farbood et al. (2013) that key-finding is essentially limited by rhythmic and melodic constraints, the actual *decoding time* for tonal processing, predicated on apprehending musical structure, was still unknown. Farbood et al. (2013) is the musical equivalent of studies that assess intelligibility of compressed speech at different rates (Dupoux & Green, 1997; Foulke & Sticht, 1969; Peele & Wingfield, 2005; Versfeld & Dreschler, 2002). The current study goes a step further and is the musical analog of Ghitza and Greenberg's (2009) study; the tempo/compression rate is not simply increased or decreased—by adding silences in a way that does not align with the natural rhythm of the sequence, we are attempting to see whether musical comprehension (in the form of key-finding) is optimized when provided additional decoding time.

A minimum decoding time for music has been hinted at in a study with a very different task and stimuli by Bigand, Poulin, Tillmann, Madurell, and D'Adamo (2003), which compared sensory versus cognitive components in harmonic priming. The stimuli for that study consisted of eight-chord sequences in which the first seven chords served as a context for a final target chord (paralleling the eight-note structure of the melodies here). They found that at 300 and 150 ms per chord, the tonal context clearly facilitated processing of the target, indicating that key-finding had successfully occurred despite the fast tempi. However, when the tempo was further increased to 75 ms per chord (13.3 Hz; 800 BPM), the effect of tonal context appeared to be overruled by sensory priming. This suggests there is a minimum amount of processing time that is necessary for key induction.

Here we address explicitly whether the perception of musical structure is subject to similar "parsing and decoding" principles hypothesized for speech and test whether the U-shaped error rate curve found for speech appears also for music. We applied the gap-insertion paradigm to time-compressed melodic sequences and asked subjects to identify the key of the melody. Potential parallels would suggest shared mechanisms between these two domains (Patel, 2003), and the study of potential oscillatory mechanisms may open up new avenues of research into basic psychoacoustic processing of music.

Method

Participants

Twenty-eight musically trained listeners participated (average age 23.64 years, *SD* = 5.73, 25 male). Formal training on a primary instrument averaged 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-rating was 3.77 (*SD* = 0.75). Average number of years of college-level music theory was 2.07 (*SD* = 1.65).

Stimulus Materials

The stimuli were based on ten melodic sequences composed by Farbood et al. (2013; Figure 1a). These melodic sequences had identical pitch content—the union of all pitches in two closely related keys differing by only one sharp or flat (e.g., C major/G major)—and ended on the same pitch, one that could be interpreted as either the tonic of or dominant of one of the two keys in question. Since the pitch sets were identical regardless of the implied key of that final note, the interpretation of that note was subject to the strong structural cues provided by the ordering of the pitches that preceded it. These structural cues included melodic intervals, the ordering of those intervals, and longer patterns of notes, delineated by contour changes, expected in typical harmonic progressions. The rationale behind using these materials as an analog to speech processing was the idea that structural cues emerge from the intervallic relationships between the pitches and are critical to key-finding much like diphones in speech combine to form syllables that lead to identification of words.

Given these melodic sequences, Ghitza and Greenberg's (2009) stimulus modification method for speech materials was then adopted. The melodies were first rendered at 1 Hz (1000 ms IOI; 60 BPM), a tempo at which key identification was highly consistent based on the results of Farbood et al. (2013), and then time-compressed by a factor of 28, a tempo at which key identification was impossible (28 Hz; 35.7 ms IOI; 1680 BPM). These compressed sequences were then altered by inserting varying durations of silences ("gaps") periodically in the audio. The unsegmented original and compressed stimuli were generated in MIDI format at the original pitch (as shown in Figure 1a) and at 22 transpositions (11 semitones up and 11 semitones down). These MIDI files were then converted to audio (WAV format) using QuickTime (grand piano timbre). The resulting audio files were segmented into consecutive audio chunks of equal duration, interspersed with gaps (Figure 1b); the final format of the stimuli was rendered in 16 kHz 16-bit mono.

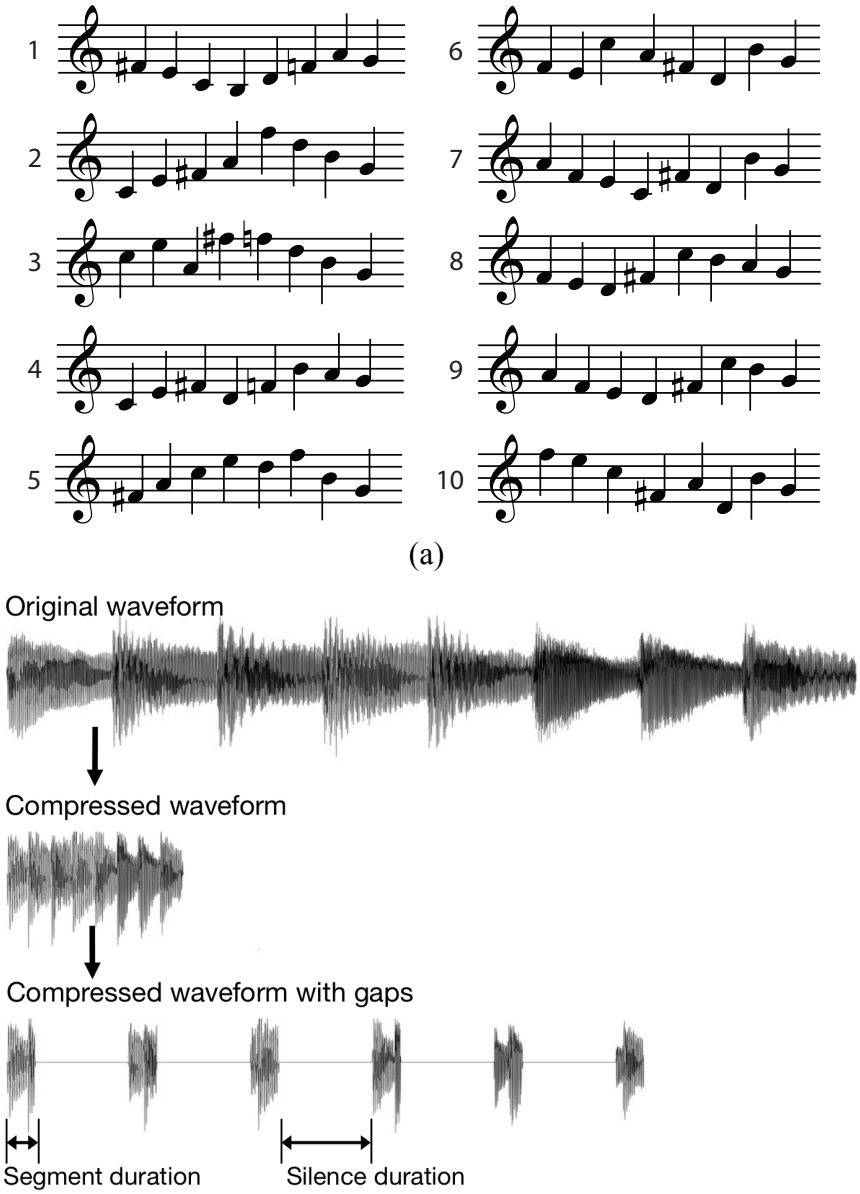

Figure 1. (a) The original melodic sequences. Sequences 1–5 are designed to sound like C major and sequences 6–10 are designed to sound like G major (before transposition). (b) Illustration of how the stimuli were created (not to scale). Top: The original melodic sequence at the original tempo. Center: The melody generated at the compressed tempo. Bottom: Compressed melody with silences inserted.

The durations of the audio segments (10, 23, 38, 55, and 65 ms) and the silence intervals (0, 40, 80, 160, 230, 640, and 1280 ms) were varied parametrically. The stimuli had a total number of audio segments that ranged from 5 (for 65 ms segments) to 29 (for 10 ms segments). The mean number of pitch fragments per segment ranged from 1.24 to 2.40. Table S1 in the Supplemental Materials provides additional information about the number of audio segments per stimulus, number of note fragments per segment, and the longest total stimulus duration for each segment length.

Procedure

Participants were seated in front of a computer and presented diotically with stimuli at a comfortable listening level over Sennheiser HD 650 headphones in a hemi-anechoic chamber. Subjects indicated whether each sequence sounded resolved (ending on an implied tonic) or unresolved (ending on an implied dominant) by entering responses into a MATLAB GUI that used Psychtoolbox extensions for audio playback (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Participants listened to 340 sequences: *twice* for each of the 10 sequences at both the uncompressed rate of 1 Hz without gaps and the compressed rate of 28 Hz also without gaps, and *once* for the 10 sequences altered at all combinations of the audio segment and silence durations (5 audio segment durations x 6 silence durations x 10 melodic sequences). Stimuli were presented in a pseudo-randomized 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, or compressed with gaps), and no stimulus was in the same key as the two preceding sequences. The transposition of each stimulus was determined based on two constraints: each sequence was transposed to least three sharps/flats away from the key of the immediately preceding trial, and transpositions alternated between upward and downward directions (i.e., above and below the pitches of the pre-transposed sequences). The experiment took approximately one hour to complete without breaks.

Results

We determined correct responses by looking at the each subject's judgments on the original, unmodified melodic sequences played at the original tempo. With the exception of Sequence 1, the judgments of key for these unmodified sequences were mostly in agreement with the expert labels from Farbood et al. (2013). Sequences 2–10 had a disagreement rate of 6.15% when compared to expert judgments, while Sequence 1 had a much higher error rate of 37.5%, indicating that this particular melody was considerably more ambiguous than the others. Unlike the case for words in speech, there is not necessarily a "correct" label for key. Although there can be nearly universal agreement, depending on the musical material in question, there commonly exists some degree of ambiguity. Thus we used each subject's own judgments to determine whether a response should be deemed correct. If a subject's responses to the unmodified versions of a particular sequence did not agree, all trials containing that sequence were removed. If the two judgments for a sequence did agree, then that judgment was interpreted as the correct response for all trials containing that sequence. After exclusion for within-subject disagreement, there remained 8500 out of 9520 total trials across subjects out of which 7500 featured stimuli with gaps. This strategy resulted in a 0% error rate for uncompressed sequences and a 44.73% error rate for compressed sequences without gaps. A chi-square goodness-of-fit test indicated that the error rate for the compressed sequences was borderline chance, $\chi^2(1, N = 500) = 3.80$, $p = .051$.

A two-way, repeated-measures analysis of variance with seven levels of silence durations (0, 40, 80, 160, 230, 640, 1280 ms) and five levels of audio segment durations (10, 23, 38, 55, 65 ms) was performed on response accuracy (percent correct for each subject); Greenhouse-Geisser corrections were used in cases where sphericity was

violated. All effects were significant at the .05 level: $F(5, 135) = 2.90$, $MSE = 199.26$, $p = .016$ for the main effect of silence duration and $F(2.66, 71.91) = 13.90$, $MSE = 351.65$, $p < .001$ for the main effect of audio segment duration. There was also a significant interaction between the two factors, $F(10.48, 283.03) = 4.18$, $MSE = 206.76$, $p < .001$, necessitating a closer look at differences across levels for each factor.

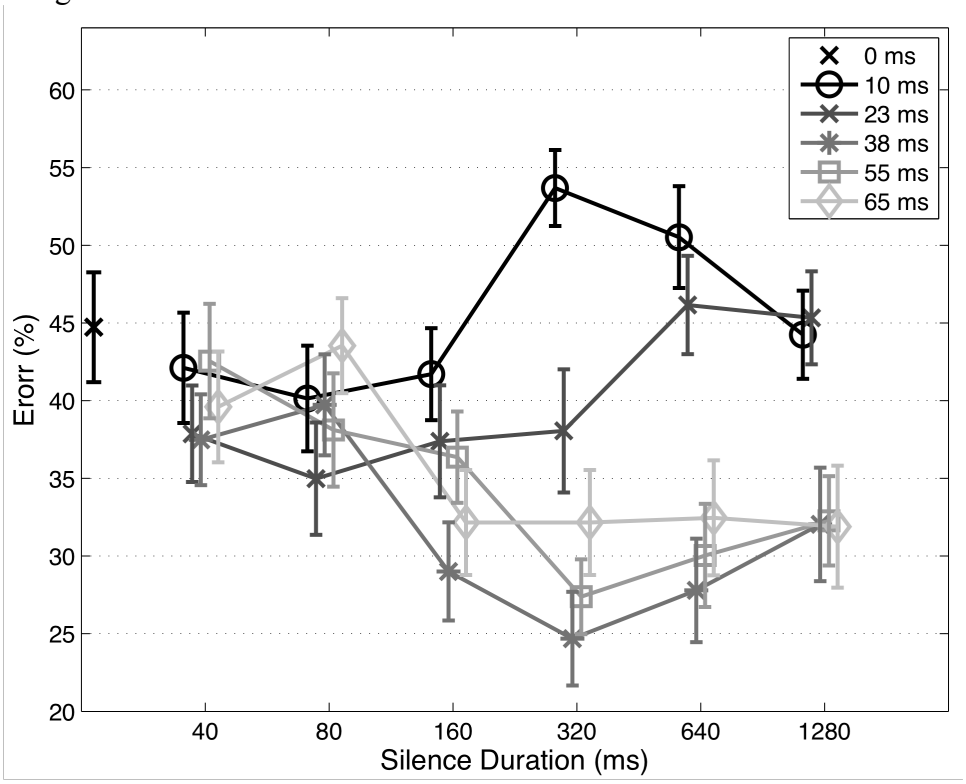

Figure 2. Mean error for all conditions graphed by audio segment duration. Error bars indicate estimated standard error.

Figure 2 shows mean error rates by audio segment duration and silence duration. Table 1 shows the results of one-way, repeated-measures ANOVAs for each silence and audio segment duration as well as post-hoc Tukey-Kramer tests. As in the case of Ghitza and Greenberg (2009), adding silences to compressed audio lowered error rate. When the data were examined by individual audio segment durations, a more complex picture emerged. The data revealed a dissociation between audio segment durations shorter and longer than the note event length. A U-shaped curve was found for the higher audio segment durations, while accuracy for shorter durations hovered at chance or even decreased in accuracy. This dissociation occurred at silence durations greater than 80 ms. At shorter silence intervals, all audio segment durations were around the near-chance error levels found for the original compressed condition without any inserted silences.

Table 1. *ANOVA results for simple effects of audio segment and silence interval.*

Factor Type	Level Type	Factor Duration (ms)	*F*	*df*	*p*	Levels with Significant Differences (ms)
Audio segment	Silence interval	10	3.25	6, 540	.0029	{40, 80, 160} & 320
		23	2.77	6, 540	.0098	None
		38	8.01	6, 540	< .001	0 & {160, 320, 640, 1280}; {40, 80} & 320; 80 & 640
		55	6.12	6, 540	< .001	0 & {320, 640, 1280}; 40 & {320, 640}
		65	4.87	6, 540	< .001	0 & {160, 320, 640, 1280}; 80 & {160, 320, 1280}
Silence interval	Audio segment	40	0.71	2.93, 395.20	.52	None
		80	1.07	4, 540	.26	None
		160	3.62	4, 540	.0059	10 & {38, 65}
		320	17.35	4, 540	< .001	10 & all; 23 & {38, 55}
		640	11.50	4, 540	< .001	{10, 23} & {38, 55, 65}
		1280	5.86	4, 540	< .001	{10, 23} & {38, 55, 65}

Note. MSE = 206.76 for all cases.

For the 10 ms audio segments, lack of adequate pitch resolution was most likely a significant factor in performance level. From a qualitative perspective, the 10 ms audio segments sounded increasingly "click-like" as the silence durations increased. There is a minimum of two to three cycles necessary for reliable pitch resolution of complex tones (Metters & Williams, 1973; Patterson, Peters, & Milroy, 1983; Pollack, 1967)—approximately 20 ms for complex tones with a fundamental frequency of 200 Hz (Ritsma & Cardozo, 1963). Given this issue of reliable pitch resolution, we added a preprocessing step in our subsequent analyses: we removed trials in which there were less than three cycles of audio for any pitches in a given sequence. This method eliminated 766 trials out of 1500 for the 10 ms case and none for any of the longer audio segments.

Assuming there is a cyclical rate at which music processing optimally occurs, we should see its signature—error rates should indicate a minimum point at the optimal rate. To further explore the preferential time window for music processing, we examined task performance by *event rate*, the musical analog of packaging rate for speech. Packaging rate for speech as defined by Ghitza and Greenberg (2009) is the periodic silence-plus-audio rate; event rate for music as defined here is the mean rate of new musical information for each stimulus. "New information" in the musical sense is pitch, which is the atomic unit necessary for interpreting musical structure and thus key. Event rate is similar to tempo, although unlike tempo, the onsets are not precisely isochronous (in the current context of artificially generated stimuli as opposed to human-performed music). We use event rate as a window into our data because the literal application of packaging rate for music does not work; packaging rate as a measure only makes sense when there is new information that occurs with each packet. This is not the case for our stimuli with 10 and 23 ms audio segments because a repeated pitch is not new information. Figure 3

shows error rate plotted by event rate. The results reveal that the error rate minimum centered around 5–7 Hz for all audio segment durations.

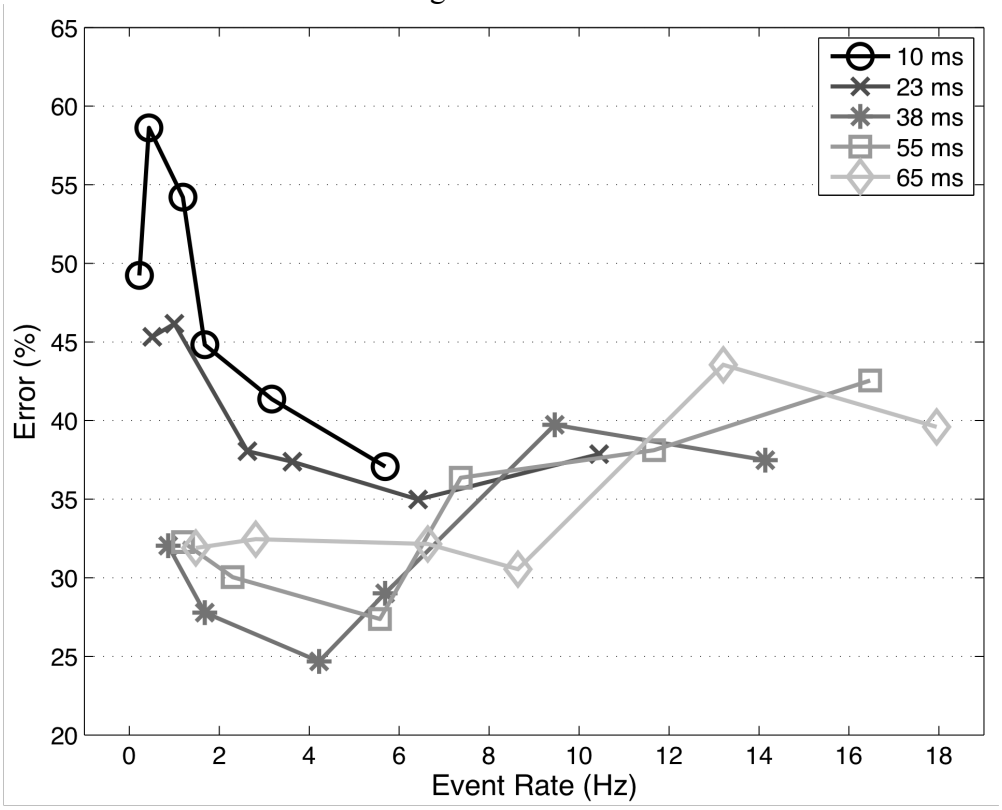

Figure 3. Mean error by event rate.

In addition to the event rate, the preferred *duty cycle* of the task was used to probe potential cyclical processing mechanisms. Duty cycle is a general engineering term used to describe the proportion of a cycle that a signal is active for periodic phenomena. Ghitza and Greenberg (2009) used it to denote the ratio of speech audio to silence. For example, no silences inserted would mean a duty cycle of 100%; a stimulus with 100 ms of audio followed by 900 ms of silence would have a duty cycle of 10%. In examining duty cycle values, we again excluded the 10 ms trials where insufficient pitch resolution was a factor. What we found was a preferential range, based on error rates, that ranged from roughly 10 to 30% (Figure 4). The results show that longer duty cycles (i.e., more audio information per unit time) does not translate into better performance, presumably because there is insufficient decoding time available.

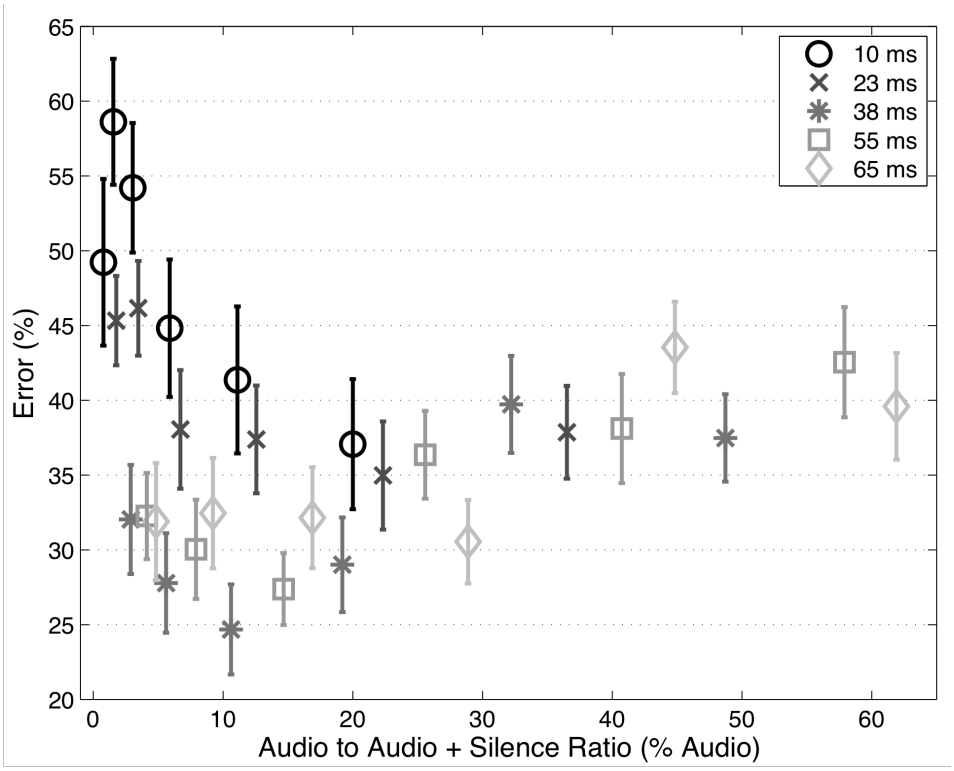

Figure 4. Mean error by duty cycle. Error bars indicate estimated standard error.

Discussion

The results of this study showed that the insertion of periodic silences between segments of compressed music significantly reduced error rate in a key-finding task. The data displayed a distinctive U-shape function when viewed by event rate, similar to results found for speech by Ghitza and Greenberg (2009). Regardless of the size of audio segments and silence intervals, the error rate minimum centered around 5–7 Hz (~140–200 ms IOI; 300–420 BPM).

Similarities and differences between the current results and Ghitza and Greenberg's (2009) findings point to the way auditory processing is tied to the specific temporal structure of its input. The observed preferential duty cycle for music processing from these results are 10–30% as opposed to 33–66% for speech. The shorter duty cycle for music might reflect the discrete-pitched nature of music; once the pitch of a note has been resolved, no further contextual information is required from the note. In contrast, syllabic envelopes are continuously evolving. At higher levels (i.e., subsequent to pitch detection), structures in music actually are processed at a slower rate than speech.

In general, the rate of change in music encompasses a wider range than in speech: the prominent range of the modulation spectrum of speech across languages tends to be 4–8 Hz/125–250 ms (e.g., Greenberg, 2006; Houtgast & Steeneken, 1985) while for melodic sequences, the ideal range is ~0.5–6.7 Hz/150–2000 ms (Farbood et al., 2013; Warren, Gardner, Brubaker, & Bashford, 1991). These modulation rates may be reflective of the natural periodicity in the neural system (cf. Buzsáki, 2006). In the case of speech, there is a correspondence between average durations of speech units and the frequency ranges of cortical oscillations (Giraud & Poeppel, 2012; Ghitza, 2011).

Phonetic features (20–80 ms), for example, are associated with low gamma and beta oscillations (15–50 Hz), while syllables and words (mean duration of 250 ms) are associated with theta (4–8 Hz) oscillations. Likewise, sequences of syllables and words embedded within a prosodic phrase (500–2000 ms) correspond to delta oscillations (1–3 Hz*).* While such results from EEG/MEG experiments are increasingly common for speech, less work has been done exploring the oscillatory nature of music processing. One such study by Carrus, Koelsch, & Bhattacharya (2011) used a frequency-based EEG approach and found that syntactic violations in chord sequences produced similar changes in delta-theta power observed after processing of syntactic violations in language. Furthermore, syntactic violations occurring at the same time for both music and speech resulted in a pattern of reduced frequency response in these bands, suggesting shared neural resources.

Extensions of Prior Work on Speech and Music

Timescale manipulations of both speech and music *without* added gaps result in U-shaped data patterns. Different mechanisms may explain the deterioration below the optimal range (time expansion) or above it (time compression). For time expansion, the limiting factor is likely the length of the working memory buffer (limit on integration); for time compression, it is the lack of decoding time and whatever other factors limit resolution. Ghitza and Greenberg (2009) focused on the time compression case in order to test the lack-of-decoding-time hypothesis, and their data exhibited a U-shaped behavior as well. Even though the results were U-shaped in both cases, the mechanisms that underlie the data pattern for the uniform timescale manipulation are distinct from those that underlie the U-shape behavior for the repackaged data. Ghitza (2011) argues that decoding time is governed by a cascade of neuronal oscillators, which guide template-matching operations at a hierarchy of temporal scales and presents a model, with a cascade of oscillators at the core, capable of emulating the counterintuitive finding of Ghitza and Greenberg (2009) data.

Farbood et al. (2013) examined the effects of timescale manipulations of music in a manner analogous to the prior work on compressed speech without gap insertions. They observed a U-shaped data pattern in a key-finding task as a function of tempo. The current study builds on and goes beyond Farbood et al. (2013) in the same way Ghitza and Greenberg (2009) departed from prior work on compressed speech—by using the repackaging procedure to examine the decoding time hypothesis for music. The ideal rates for the key-finding task in Farbood et al.'s (2013) study encompassed a large plateau (0.5–6.7 Hz; 150–2000 ms IOI; 30–400 BPM) where performance was at ceiling. This aligns well with current estimates of the peak spectrum of music, which is around 2–3 Hz for a wide range of musical pieces (Ding, Patel, & Poeppel, in review). If oscillations form the neuronal basis for this perceptual analysis, they would be in the delta range (~1–3 Hz). In the present results, however, the current data implicate a higher modulation rate, and by extension, oscillation. The best performance lies in the 5–7 Hz range, associated with theta (4–8 Hz) activity. There is, as a consequence, a tension in the old and new results about a possible oscillatory interpretation. The current data suggest that parsing or chunking an input stream at roughly 5 Hz holds for both speech and music. The older data, in contrast, underscore a potential difference between domains. If oscillatory neuronal activity plays a central role, the data to date cannot adjudicate

between the alternatives.

The similarities between the findings for music and speech using this silence-insertion paradigm provide compelling clues into possible oscillatory mechanisms in the auditory domain. Nonetheless, much remains to be learned about these processes. Given these results, as well as evidence suggesting shared neural resources between syntactic processing of music and speech (Fedorenko, Patel, Casasanto, Winawer, & Gibson, 2009; Koelsch, Gunter, Wittfoth, & Sammler, 2005; Patel, 2003), Western tonal music provides an ideal medium for the comparative exploration of these mechanisms.

References

Bigand, E., Poulin, B., Tillmann, B., Madurell, F., & D'Adamo, D. A. (2003). Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 159–171. doi:10.1037/0096-1523.29.1.159

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Buzsáki, G. (2006). *Rhythms of the Brain*. New York: Oxford University Press.

Carrus, E., Koelsch, S., & Bhattacharya, J. (2011). Shadows of music-language interaction on low frequency brain oscillatory patterns. *Brain and Language*, *119*(1), 50–57. doi:10.1016/j.bandl.2011.05.009

Ding, N., Patel, A. D., & Poeppel, D. (under review). Temporal modulations in speech and music.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, *95*(2), 1053–1064. doi:10.1121/1.408467

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 914–927.

Farbood, M. M., Marcus, G., & Poeppel, D. (2013). Temporal dynamics and the identification of musical key. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 911–918. doi:10.1037/a0031087

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition*, *37*(1), 1–9. doi:10.3758/MC.37.1.1

Foulke, E., & Sticht, T. G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, *72*(1), 50-62.

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*, 1–13. doi:10.3389/fpsyg.2011.00130

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Frontiers in Psychology*, 1–12. doi:10.3389/fpsyg.2012.00238

Ghitza, O. (2013). The theta-syllable: a unit of speech information defined by cortical function, 1–5. doi:10.3389/fpsyg.2013.00138

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113–126. doi:10.1159/000208934

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. doi:doi:10.1038/nn.3063

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (2nd ed.). Hoboken, NJ: Wiley.

Greenberg, S. (2006). A Multi-tier framework for understanding spoken language. In S. Greenberg & W. Ainsworth (Eds.), *Listening to Speech: An Auditory Perspective* (pp. 411–433). Mahwah, NJ: Erlbaum.

Houtgast, T., & Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, *77*(3), 1069–1077. doi:10.1121/1.392224

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(ECVP Abstract Supplement).

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*(10), 1565–1577.

London, J. (2012). *Hearing in Time: Psychological Aspects of Musical Meter* (2nd ed.). New York: Oxford University Press.

Patel, A. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7).

Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(6), 1315–1330. doi:10.1037/0096-1523.31.6.1315

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time." *Speech Communication*, *41*(1), 245–255. doi:10.1016/S0167-6393(02)00107-3

Ritsma, R. J., & Cardozo, B. L. (1963/64). The perception of pitch. *Philips Technical Review*, *64*(2/3), 37–43.

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *336*(1278), 367–373.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304. doi:10.1126/science.270.5234.303

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge: MIT Press.

Stevens, K. N. (2005). Features in speech perception and lexical access. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 125–155). Blackwell.

Versfeld, N. J., & Dreschler, W. A. (2002). The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, *111*, 404–408.

Warren, R. M., Gardner, D. A., Brubaker, B. S., & Bashford, J. A. (1991). Melodic and nonmelodic sequences of tones: Effects of duration on perception. *Music Perception*, *8*, 277–289.