Determining the end of a musical turn

Citation for published version:
Hadley, L, Sturt, P, Moran, N & Pickering, M 2018, 'Determining the end of a musical turn: Effects of tonal cues', *Acta Psychologica*, vol. 182, pp. 189-193. https://doi.org/10.1016/j.actpsy.2017.11.001

Digital Object Identifier (DOI):
10.1016/j.actpsy.2017.11.001

Published In:
Acta Psychologica

Determining the end of a musical turn: Effects of tonal cues

Lauren V. Hadley, Patrick Sturt, Nikki Moran, and Martin J. Pickering

University of Edinburgh

Running Head: Determining the end of a musical turn

Author Note

Lauren V. Hadley, Department of Psychology, University of Edinburgh; Patrick Sturt, Department of Psychology, University of Edinburgh; Nikki Moran, Reid School or Music, University of Edinburgh; Martin J. Pickering, Department of Psychology, University of Edinburgh.

Address for correspondence

Lauren Victoria Hadley

Department of Psychology

University of Edinburgh

Edinburgh EH8 9JZ

E-mail: lauren.hadley@cantab.net

phone: +44 (0) 131 650 3391

Acknowledgements

L.V. Hadley was supported by a grant from the Economic and Social Research Council.

ABSTRACT

Successful duetting requires that musicians coordinate their performance with their partners. In the case of turn-taking in improvised performance they need to be able to predict their partner's turn-end in order to accurately time their own entries. Here we investigate the cues used for accurate turn-end prediction in musical improvisations, focusing on the role of tonal structure. In a response-time task, participants more accurately determined the endings of (tonal) jazz than (non-tonal) free improvisation turns. Moreover, for the jazz improvisations, removing low frequency information (<2100Hz) - and hence obscuring the pitch relationships conveying tonality - reduced response accuracy, but removing high frequency information (>2100Hz) had no effect. Neither form of filtering affected response accuracy in the free improvisation condition. We therefore argue that tonal cues aided prediction accuracy for the jazz improvisations compared to the free improvisations. We compare our results with those from related speech research (de Ruiter et al., 2006), to draw comparisons between the structural function of tonality and linguistic syntax.

Keywords: Turn-taking, musical coordination, tonality, prediction, improvisation

INTRODUCTION

Accurate temporal coordination between members of a musical ensemble is essential for coherent performance (Keller, 2008), and such coordination requires performers to predict each other's behaviour (Pecenka & Keller, 2011; Phillips-Silver & Keller, 2012). The need to predict each other is particularly apparent in turn-taking contexts, as turns must be accurately timed between individuals when the primary communicative role switches. Musicians have access to auditory cues such as pitch, duration, and intensity, as well as visual cues such as gaze or body movement, to help them determine the end of a co-performer's turn. As the audio information is the primary musical signal, and visual information can be obscured in performance contexts (because of positioning, lighting etc.), we focus on auditory cues. These auditory cues are relevant both for performers predicting turn-ends when playing with a partner, and also for non-performers predicting the music that they passively listen to. We therefore had listeners predict turn-ends for jazz and free improvisations under various conditions to determine the importance of cues relating to tonality in musical prediction.

Accurate turn-end prediction is critical not only for music performance but also for linguistic conversation (Stivers et al., 2009). In fact, research on conversation has investigated the importance of a variety of auditory cues. De Ruiter et al. (2006) presented listeners with recorded utterances in their original form, lowpass filtered, or with a flattened contour. Lowpass filtering removed high frequencies commonly used to distinguish consonants, obscuring structural word-level cues including semantic and syntactic information. Flattening the contour set the pitch contour to horizontal (i.e. monotone), removing intonation cues. While listening, participants were told to 'press a button in front of them at the moment they thought the speaker would be finished speaking (Dutch: *is uitgesproken*).' The authors argued that 'The instruction encouraged the subjects to try to ANTICIPATE this moment, and not wait until the fragment stopped playing' (de Ruiter et al., 2006, p. 523). Listeners were able to predict the end of a conversational turn equally well when it was presented in its original form (M = 186ms prior to offset) or when its contour was flattened (M ~ 160ms prior to offset [estimated from figure]), but accuracy dropped when words were obscured through lowpass filtering (M ~ 470ms prior to offset [estimated from figure]). This suggests that information conveyed in the upper frequencies of speech, including semantic and syntactic information, is vital for accurately predicting turn-ends of speech.

In the current paper, we will investigate the accuracy with which listeners predict the end of turns in a musical improvisation: a form of performance in which musicians generate coherent musical utterances in real time, without recourse to a pre-planned script. While certain aspects of an improvisation may nonetheless be predetermined (for example the harmonic progression of a jazz standard or an individually generated plan for an upcoming utterance), many decisions regarding content are made in real time (Berkowitz, 2010; Ashley, 2016). We investigate how accuracy differs between musical genres (jazz and free improvisation), as well as the structural cues for prediction used within these genres, by evaluating the importance of high and low frequency information.

In music each note is a periodic waveform, and the number of times that this waveform repeats per second is its fundamental frequency. The fundamental frequency is the note's lowest frequency component, and is commonly understood as its pitch. However, each note's spectrum also includes multiples of this fundamental frequency, which make up the harmonics. Although it is possible to perceive pitch from these harmonics alone, inferring pitch without the fundamental frequency takes substantially longer than when the fundamental is present (Winkler et al., 1997). Pitch provides structural information, with pitches organised into a hierarchy (in terms of their stability and probability) whose pattern may be perceived as tonality. Tonality thus describes the probability of different note-to-note progressions (and by extension, larger-scale progressions). Listeners do not require musical training beyond normal exposure to perceive tonal structure during music listening (see Huron, 2006 for a broad summary), and recognise anomalous progressions (Koelsch et al., 2000; Pearce et al., 2010). The lower frequencies of musical notes, therefore, convey their pitch and are critical for conveying tonality, which is commonly compared with syntax and semantics in language (Koelsch, 2005; Patel, 2003; Steinbeis & Koelsch, 2008).

We investigated turn-end prediction for two forms of musical improvisation, namely jazz and free, across a range of listeners using a paradigm similar to De Ruiter et al. (2006). Broadly speaking, jazz improvisation uses a tonal framework to constrain both note-to-note and larger-scale structure

(Barrett & Peplowski, 1998), whereas free improvisation does not use a tonal framework, but uses clustering to constrain large-scale structure (Dean, Bailes, & Drummond, 2014). We first compared turn-end accuracy for original jazz and free improvisations. We then compared turn-end accuracy for jazz and free improvisations that had been highpass filtered, obscuring pitch information (and hence in the jazz improvisations, tonal information). Such filtering has been used to investigate the salience of specific spectral cues in music and speech research (Moore & Tan, 2003; Schellenberg, Iverson, & Mckinnon, 1999). We expected listeners' turn-end accuracy to be higher for the (tonal) jazz than the (non-tonal) free improvisations. We further hypothesised that if tonal cues are critical for predicting the end of a musical turn, removing low frequencies should impair accuracy for jazz improvisations more than free improvisations. We included a control manipulation (lowpass filtering) to ensure that effects were not due to general spectral depletion.

METHODS

Participants

Forty-seven participants took part in the study.[1] We excluded participants if they responded to less than 70% of stimuli (1 participant), or produced responses that fell more than 2.5SD from the participants' overall mean (1 participant). Of the remaining 45 participants, 29 were female, and ages ranged from 18 to 65 (M = 27, SD = 11). We recruited participants by email, either within the university or through musician networks, to ensure a range of musical training. A standard musical questionnaire was used to record extent of musical training on a 7-point scale (0 / 0.5/ 1 / 2 / 3-5 / 6-9 / 10 or more years; Müllensiefen, Gingras, Stewart, & Musil, 2011), revealing that participants had a broad range of musical training experience. Nine participants had no training, one had 1 year, four had 2 years, 11 had 3-5 years, 11 had 6-9 years, and nine had 10 or more years.

[1] As there was no way to estimate effect size a priori, sample size was determined through reference to studies using similar paradigms in language (cf. Magyari & de Ruiter, 2012)

Materials

The experimental items were extracted from the recordings by Moran and Keller (2016), which included idiomatic duets from six pairs of jazz improvisers and six pairs of free improvisers. The jazz improvisers followed the harmonic progression of a jazz standard (*Autumn Leaves*, J. Kosma, 1945), and played with a regular pulse in the range of 400-500ms (equivalent to a range of 120-150 beats per minute). Meanwhile, the material generated by the six pairs of free improvisers was characterised by the absence of consistent and predictable tonal structure, and the absence of a regular and reliable pulse (and hence metre). All improvisers specialised in the style they performed and had at least 7 years of performance experience. The final phrases of 90 solo improvisation turns within these recordings were determined (47 jazz, 43 free) with phrases defined as perceptually complete musical units as judged by the first author (a music graduate). These phrases were taken from a point of silence until the end of the final note (endpoint determined in Logic Pro), and did not include any sounds other than those of the instrument.

The maximum f0 in these stimuli was 1846Hz and the mean was 336Hz. Ritsma (1967) found that frequencies of 500-1500Hz are most important to determine the pitch of complex tones with an f0 of up to 400Hz, and hence we generated a *highpass* condition excluding all f0 frequencies (threshold 2100Hz), and a *lowpass* condition incorporating all f0 frequencies (threshold 2100Hz). Highpass filtering obscured the pitch of the jazz and free improvisations, and, as tonality is based on pitch, the tonal framework of the jazz improvisations. Lowpass filtering, on the other hand, did not affect the pitch of the jazz and free improvisations but provided a spectrally reduced control. We additionally included the *normal* condition (original recordings). Stimuli in each condition were equated in root mean square amplitude to level the extracts in loudness.

Each participant heard one version of each item. Items were divided into three groups (approximately matched in distribution of improvisation styles, durations, and instruments), and each participant heard one group in each condition (i.e., 30 normal, 30 highpass, and 30 lowpass stimuli), in an individually randomised order (see Table 1).

Table 1

Item groups for experimental stimuli

		Group 1	Group 2	Group 3
Improvisation Style	Free	15	14	14
	Jazz	15	16	16
Instruments	Saxophone	13	13	14
	Piano	5	6	5
	Drum	3	3	3
	Guitar	2	2	2
	Clarinet	3	2	2
	Violin	1	1	2
	Cello	1	1	1
	Trumpet	1	2	1
	Bass guitar	1	0	0
Total	Mean Duration	6.3	6.5	6.4
	Min Duration	2.6	2.0	1.9
	Max Duration	10.7	10.7	9.8

Procedure

Each participant was randomly assigned to one of three stimulus lists (15 participants per list), and instructed: 'Press the button at the moment you think the musician will have finished their solo. Try to anticipate the end of their turn.' Participants were warned that some stimuli would sound unusual, but to give their best guess as to when the solo would be finished. The experiment began with a practice block, in which one example of each spectral manipulation was presented, followed by the main experiment. On each trial, the participant pressed a button to indicate readiness, after which a fixation cross was presented for 500ms, followed by the stimulus. Participants responded by pressing the button again, which cut the music and ended the trial. (The music did not continue after the button press, as we judged that participants might then wait for the end of the performances before responding.) If participants did not press the button during a trial, a time out occurred 2 seconds after the end of the stimulus. The procedure then repeated. Participants were told that they could take breaks between trials. Button press timings were recorded relative to the start of the stimulus. The lack of a button press was recorded as a missed trial.

Design and Data Analysis

The experiment involved a 2 Improvisation (free, jazz) x 3 Spectral Manipulation (normal, highpass, lowpass) design. After exploring the distribution of responses, we analysed response accuracy. During joint music performance, both early and late entries may be considered equally erroneous. Our instructions emphasised prediction, but specified that this prediction should lead to responses occurring at the stimulus offset. We therefore defined response accuracy as temporal proximity to the stimulus offset, and treated early and late responses as equivalently inaccurate because any deviation from the offset constitutes inaccurate prediction. Any trials in which a participant did not respond, or in which responses were more than 2.5SD from the participant's mean, were removed.[2] This led to the removal of 7.21% of trials.

[2] All analyses show the same main effects and interactions whether we removed outliers more than 2sd, 2.5sd, or 3sd from the participant mean, or 2sd, 2.5sd, or 3sd from the item mean.

RESULTS

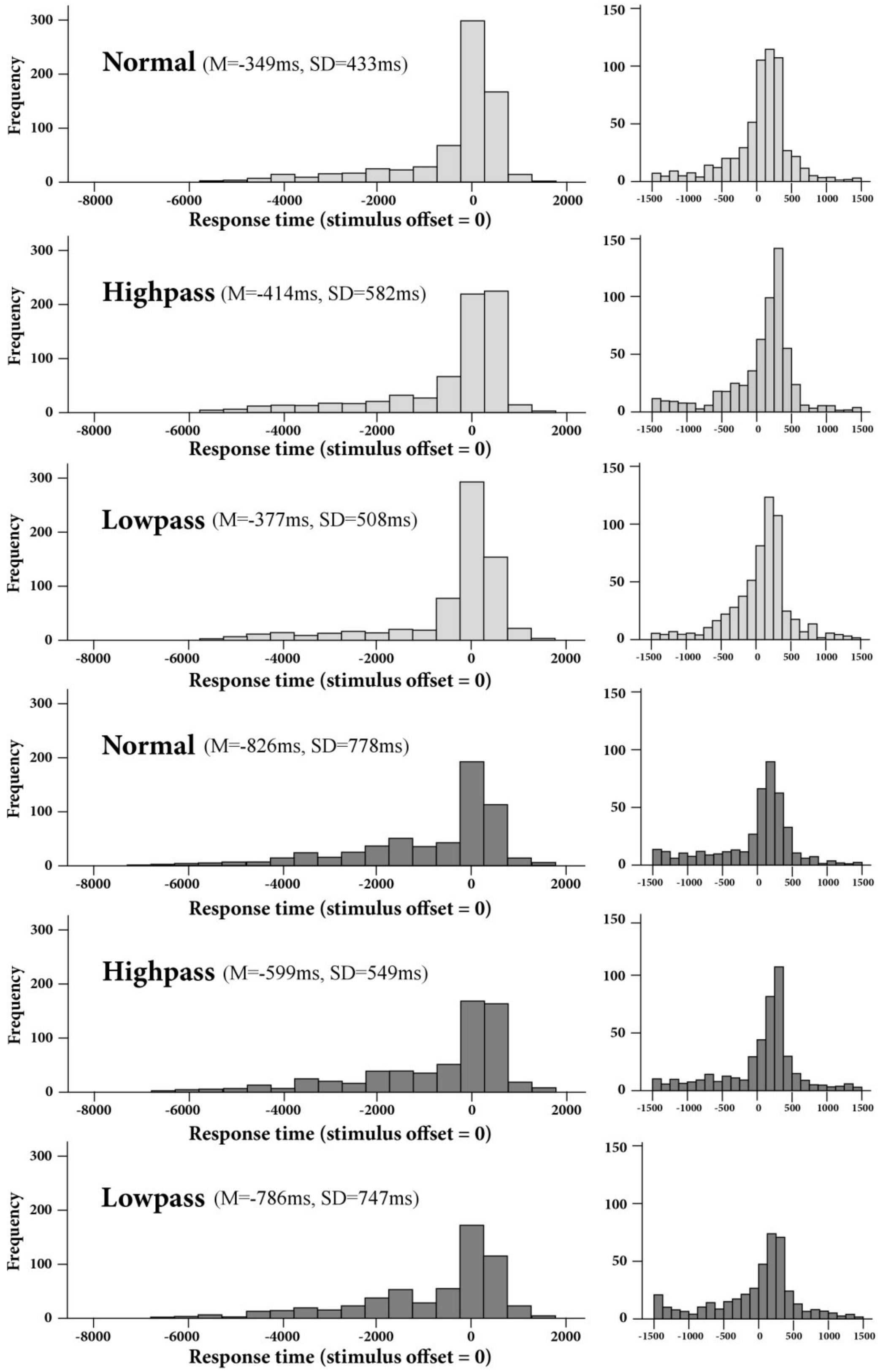

Figure 1. Left: Distributions by condition across the entire phrase (500ms bins). Right: Distributions by condition around offset (125ms bins). Light = Jazz improvisation, Dark = Free improvisation.

Table 2

Absolute response accuracy by participant (ms)

Condition	Mean	SD
Normal Jazz	357	178
Highpass Jazz	459	266
Lowpass Jazz	378	194
Normal Free	694	375
Highpass Free	623	327
Lowpass Free	700	329

For expository purposes, descriptive statistics are shown without the log transform

To investigate participants' response accuracy we log-transformed the data as a correction for skew. Absolute response accuracy (derived from the log-transformed data) is presented in Table 2. Participants were more accurate at judging the end of the jazz than the free improvisations ($F_1(1, 44) = 108.781$, $p < .001$, partial $\eta^2 = 0.712$; $F_2(1,88) = 16.404$, $p < .001$, partial $\eta^2 = 0.157$). There was no main effect of spectral manipulation ($F_1(2, 88) = 1.380$, $p = .257$; $F_2(2, 176) = 0.878$, $p = .418$), but there was an interaction between improvisation type and spectral manipulation ($F_1(2, 88) = 7.206$, $p = .001$, partial $\eta^2 = 0.141$; $F_2(2, 176) = 3.102$, $p = .047$, partial $\eta^2 = 0.034$). For the jazz improvisations, accuracy was lower for the highpass than the normal condition ($t_1(44) = -3.297$, $p = .002$, d = 0.491; $t_2(46) = -2.936$, $p = .005$, d = 0.428), but there was no difference between the lowpass and normal conditions ($t_1(44) = -0.727$, $p = .471$; $t_2(46) = -0.814$, $p = .420$). In the free improvisations there were no reliable accuracy differences between conditions (normal vs highpass $t_1(44) = 1.228$, $p = .226$; $t_2(42) = 0.624$, $p = .536$; normal vs lowpass $t_1(44) = -0.563$, $p = .576$; $t_2(42) = -0.481$, $p = .633$). See Figure 2.

Correlations between musical training level and response accuracy was not significant for any condition ($ps > .05$).

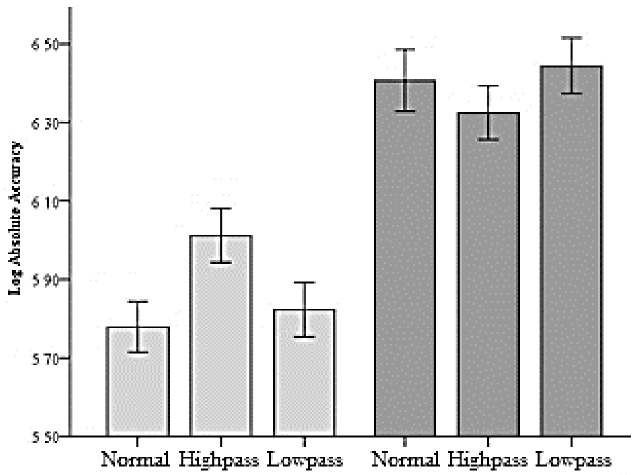

Figure 2. Absolute response accuracy (log-transformed) by participant. Error bars show standard error of the mean. Light = Jazz improvisation, Dark = Free improvisation.

DISCUSSION

We have shown that listeners are able to predict the end of a jazz improvisation turn more accurately than the end of a free improvisation turn, and that prediction accuracy does not correlate with musical training. The high level of accuracy achieved in the jazz condition suggests that musical prediction ability is highly developed in the general population, and can be used explicitly when required. Furthermore, we have shown that highpass filtering impairs turn-end prediction for jazz improvisations but not free improvisations. The detrimental effect of highpass filtering in jazz but not free improvisation indicates that information in the lower frequencies is specifically used to aid prediction accuracy in the jazz condition, suggesting that tonal information facilitates prediction. Furthermore, a lowpass filter condition had no effect on turn-end prediction, indicating that the accuracy decrease for highpass jazz (compared to normal jazz) was not a general effect of spectral depletion.

It is striking that participants across a range of musical skill were able to make such accurate predictions about the ending of an improvised musical turn lasting from 3s to 11s. This paradigm demonstrates the ability of listeners to anticipate and recognise the end of a musical turn with high temporal precision. Comparing our results with those of de Ruiter et al (2006), we see similar response timings and distributions across music and language. De Ruiter et al. found that responses occurred 186ms on average before the end of an unmanipulated utterance (slightly less than an average Dutch syllable, Eefing, 1991), and we showed that responses occurred on average 349ms before the end of an unmanipulated jazz phrase (slightly less than the average beat in these stimuli). Such pre-emptive responses to the jazz stimuli are perhaps unsurprising, as the confirmation of a tonal ending may occur several beats before the final note offset. Furthermore, it is possbile that listeners of both tonal music and spoken language are able to determine the end of a turn with an accuracy of slightly under one unit (syllable or beat).

We suggested that tonal information may be used similarly to word-level (syntactic or semantic) information in language for turn-end prediction. We showed that while listeners were more accurate at predicting (tonal) jazz improvisations than (non-tonal) free improvisations, this accuracy decreased when jazz improvisations were highpass filtered (and tonal information obscured). This is consistent with tonal information being used to predict musical turn ends. However, we acknowledge an alternative explanation: that the regular metric frameworks inherent in the jazz (but not the free) stimuli facilitated prediction of turn ends in the jazz condition. The perception of metre arises from a regular beat structure in combination with pitch-based factors such as tonal strength, harmonic structure, or melodic contour (London, 2012). Beat perception is based on information about note onset and regularity that would not be obscured by highpass filtering (Rodet & Jaillet, 2001, Bello et al., 2005), and it is unlikely that impaired beat perception drove our effects. However, it is nonetheless possible that by obscuring pitch and/or tonality in the jazz stimuli, highpass filtering also reduced metric salience.

There remains a question of which cues may be used for prediction in free improvisation. Musicians indeed seem able to take turns in such music, implying that it is possible to make predictions about when a turn is about to end (Moran et al., 2015; MacDonald et al., 2012). We suggest that in such music, which does not use a clear auditory structural framework (such as tonality or metre), visual cues such as body movement may become critical. This speculation is supported by the study of Moran et al. (2015), in which participants successfully identified authentic pairs of free improvisers from point-light audio-visual displays of the performers' body movement, but were not able to reliably identify equivalent displays of authentic standard jazz improvisers. An alternative explanation, however, relates to the pervasiveness of jazz improvisation compared to free improvisation in everyday life. Although even non-musicians are known to internalise standard tonal structure such as that used in jazz improvisation through passive enculturation (Hannon & Trainor, 2007), free improvisation uses different, genre-specific, structural techniques that are unlikely to be similarly internalised (since free improvisation is less ubiquitous). Thus a lack of familiarity with the type of pitch cues that convey structure may have contributed to the undifferentiated performance across spectral manipulations in the free condition.

We have shown than listeners with a range of musical training are able to explicitly predict the end of a musical turn relatively accurately. Our findings suggest that their predictions make use of tonal constraints in the performed musical material, and furthermore that these predictions have a mean accuracy of less than one musical beat. This accuracy is comparable to that for linguistic utterances, and hence we suggest that tonal information in music provides cues for prediction comparable to semantic and syntactic word-level information in language. Finally, although we examined turn-end prediction in a purely listening context, we expect that these findings would also be relevant to musicians in a performance context. In jazz improvisation, then, we suggest that tonal cues would be used by the non-soloing musician to predict when a partner is likely to end and thus begin preparing their own entry, while in free improvisation, we speculate that alternative (possibly visual) cues would provide the structural information essential for turn-end prediction.

REFERENCES

Ashley, R. (2016). Musical improvisation. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (2nd ed., pp. 667–679). Oxford: Oxford University Press.

Barrett, F., & Peplowski, K. (1998). Minimal structures within a song: An analysis of “All of Me.” *Organization Science*, *9*(5), 558–560.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, *13*(5), 1035-1047.

Berkowitz, A. (2010). *The improvising mind: Cognition and creativity in the musical moment*. Oxford: Oxford University Press.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

De Ruiter, J.-P., Mitterer, H., & Enfield, N. J. (2006). Projecting the End of a Speaker’s Turn: A Cognitive Cornerstone of Conversation. *Language*, *82*(3), 515–535.

Dean, R., Bailes, F., & Drummond, J. (2014). Generative Structures in Improvisation: Computational Segmentation of Keyboard Performances. *Journal of New Music Research*, *43*(2), 224–236.

Eefting, W. (1991). The effect of “‘information value’” and “‘accentuation’” on the duration of Dutch words, syllables, and segments. *The Journal of the Acoustical Society of America*, *89*(1), 412.

Hannon, E. E., & Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends in Cognitive Sciences*, *11*(11), 466–72.

Huron, D. B. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, Mass.: MIT press.

Keller, P. (2008). Joint action in music performance. In G. Morganti, A. Carassa, & G. Riva (Eds.), *Enacting Intersubjectivity: A Cognitive and Social Perspective to the Study of Interactions* (pp. 205–221). IOS Press, Amsterdam.

Koelsch, S. (2005). Neural substrates of processing syntax and semantics in music. *Current Opinion in Neurobiology*, *15*(2), 207–12.

Koelsch, S., Gunter, T., Friederici, A., & Schroger, E. (2000). Brain indices of music processing:“nonmusicians” are musical. *Journal of Cognitive Neuroscience*, *12*(3), 520–541.

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*(38), 15443–15448.

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

MacDonald, R., Wilson, G., & Miell, D. (2012). Improvisation as a creative process within contemporary music. *Musical Imaginations: Multidisciplinary perspectives on creativity, performance and perception*, 242-255.

Magyari, L., & de Ruiter, J. P. (2012). Prediction of Turn-Ends Based on Anticipation of Upcoming Words. *Frontiers in Psychology*, *3*, 376.

Moore, B., & Tan, C. (2003). Perceived naturalness of spectrally distorted speech and music. *The Journal of the Acoustical Society of America*, *114*(1), 408–419.

Moran, N., Hadley, L., Bader, M., & Keller, P. (2015). Perception of "Back-Channeling"Nonverbal Feedback in Musical Duo Improvisation. *PloS One*, *10*(6), e0130070.

Moran, N., & Keller, P. (2016). Improvising Duos - Audio only. *[dataset]*, University of Edinburgh, Edinburgh College of Art, http://dx.doi.org/10.7488/ds/1685.

Müllensiefen, D., Gingras, B., Stewart, L., & Musil, J. (2011). The Goldsmiths Musical Sophistication Index (Gold-MSI): Technical Report and Documentation v0. 9. *London: Goldsmiths, University of London.*

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–81.

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302-313.

Pecenka, N., & Keller, P. E. (2011). The role of temporal prediction abilities in interpersonal sensorimotor synchronization. *Experimental Brain Research*, *211*(3-4), 505–15.

Phillips-Silver, J., & Keller, P. E. (2012). Searching for roots of entrainment and joint action in early musical interactions. *Frontiers in Human Neuroscience*, *6*, 26.

Ritsma, R. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *The Journal of the Acoustical Society of America*, *42*(1), 191–198.

Rodet, X., & Jaillet, F. (2001). Detection and modeling of fast attack transients. In *ICMC* (pp. 1-1).

Schellenberg, E., Iverson, P., & Mckinnon, M. (1999). Name that tune: Identifying popular recordings from brief excerpts. *Psychonomic Bulletin & Review*, *6*(4), 641–646.

Steinbeis, N., & Koelsch, S. (2008). Shared neural resources between music and language indicate semantic processing of musical tension-resolution patterns. *Cerebral Cortex*, *18*(5), 1169–78.

Winkler, I., Tervaniemi, M., & Näätänen, R. (1997). Two separate codes for missing-fundamental pitch in the human auditory cortex. *The Journal of the Acoustical Society of America*, *102*(2), 1072-1082.