Music intervals in dialogic speech: Psychological disposition modulates distance accuracy among interlocutors' nonlocal F0 emissions in real-time dyadic conversation

Juan-Pablo Robledo del Canto, Esteban Hurtado, Felipe Prado, Domingo Román, Carlos Cornejo

Original Article

Music intervals in dialogic speech: Psychological disposition modulates distance accuracy among interlocutors' nonlocal F0 emissions in real-time dyadic conversation

Juan P Robledo[1,4], Esteban Hurtado[1,2], Felipe Prado[1], Domingo Román[3] and Carlos Cornejo[1]

---

[1] Language, Interaction and Phenomenology Laboratory (LIF), Psychology Department, Pontificia Universidad Católica de Chile, Santiago, Chile
[2] Laboratory of Cognitive and Social Neuroscience, UDP-INECO Foundation Core on Neuroscience (UIFCoN), Universidad Diego Portales, Santiago, Chile
[3] Phonetics Laboratory, Humanities Faculty, Pontificia Universidad Católica de Chile, Santiago, Chile
[4] Centre for Music and Science, Faculty of Music, University of Cambridge, Cambridge, United Kingdom

Corresponding Author:

Juan Pablo Robledo, Wolfson College, Barton Road, Cambridge, UK, CB3 9BB

Email: jper2@cam.ac.uk

Abstract

Drawing on the notion of musical intervals, recent studies have demonstrated the presence of frequency ratios within human vocalisation. Methodologically, these studies have addressed human vocalisation at a single-individual level. In the present study, we asked whether patterns such as musical intervals are also detected among the voices of people engaging in a conversation as an emerging interpersonal phenomenon. A total of 56 participants were randomly paired and assigned to either a control or a low-trust condition. Frequency ratios were generated by juxtaposing nonlocal fundamental frequency (F0) emissions from two people engaged in each individual dyadic conversation. Differences were found among conditions, both in terms of interval distribution and accuracy. This result supports the idea that psychological dispositions modulate the musical intervals generated between participants through mutual real-time vocal accommodation. These results underscore the socio-intentional dimension of music in vocal pitch interplay.

Keywords

Music cognition, dialogic speech, vocal prosody, musical intervals, nonlocal dependencies, , trust

Introduction

Recent studies have described the presence of frequency ratios within human vocalisation by drawing on the notion of musical intervals. Some studies have explored physiological properties of the average human vocal apparatus and the periodic sounds that it physically generates (Schwartz, Howe, & Purves 2003; Ross, Choi & Purves, 2007). Psychophysiological factors have also been investigated by relating different affective states to features of the human voice (Bowling, Gill, Choi, Prinz, and Purves, 2010; Bowling, Sundararajan, Han, & Purves, 2012). Furthermore, particular emotions have shown associations with vocally generated musical intervals (Curtis and Barucha, 2010). These results notwithstanding, most of these studies have addressed human vocalisation as an individual phenomenon. Because human oral communication is usually generated in reference to someone else, it is reasonable to ask whether there are patterns of musical intervals among people while they sustain a conversation. In the present study, this question was explored by assessing the interpersonal psychophysiological levels of musical intervals in dialogical speech.

Music in speech

Dating back to Darwin (1871), important connections have been observed at a scientific level between human vocal communication and musical phenomena,

suggesting that musical components of speech have influenced the evolution of human communication (Brown, 2000; Cross, & Woodruff, 2009). At an ontogenetic level, researchers have investigated musical features in infant-caregiver interactions (Trehub, 2003; Trainor & Desjardins, 2002; Trainor, Austin, & Desjardins, 2000). In this context, *infant-directed speech* (IDS) has been described as a communicative device with specific acoustic features (exaggerated pitch contours, larger pitch range, and slower tempo) that resembles music rather than language (Fernald, 1992). Trainor et al. (2000) have argued that the same features can be found in *adult-directed speech* (ADS) whenever people express emotion, being otherwise not noticeable due to an inhibitory process that is typically present in adult-adult interactions. Malloch and Trevarthen (2009) have also emphasised that IDS's musical features are present in interactions among adults in the form of discrete vocal events (Malloch & Trevarthen, 2009). Many dimensions of meaning can be not only situated but also conveyed throughout phonetic means (Cruttenden, 1997). Among these dimensions, the 'attitudinal' or 'affective stance' (Ochs, 1996) consists of how speakers convey their attitudes toward what is being said or the person to whom it is said. Trainor et al. (2000) have proposed that the introduction of language in infants corresponds to the end of prosody as the exclusive vocal vehicle for conveying meaning. This

transition would mark the beginning of a progressive social restraint of the prosodic expression of emotion. Such restraint, they hypothesised, would allow more cognitive and reflective reactions to prevail over immediate emotional ones.

Pitch and musical intervals

Most of the work on musical aspects of vocal pitch describes it in terms of melodic contour (Fernald, 1989; Fernald, 1991; Stern, Spieker & MacKain, 1982; Trehub, Trainor & Unyk, 1993; Malloch & Trevarthen, 2009). A limitation of this type of analysis, however, is that it describes pitch behaviour in a rather global way – in terms of, e.g., 'large', 'rising', 'lower' contours – overlooking more accurate relations between frequencies. To assess pitch and pitch relations in vocal research, the notion of musical intervals (Table 1) seems to provide advantage, considering they have been the main means for accurately addressing pitch distances (Burkholder, Grout, & Palisca, 2010). In fact, there is evidence that musical intervals have an impact at a psychological level from an early age and throughout life (Masataka, 2006; Trainor, Christine, Tsang & Cheung, 2002; Schellenberg, & Trehub, 1999; Smith, & Williams, 1999; Oelmann, & Laeng, 2008). Based on this evidence, recent studies have tested the role of musical intervals in speech.

Although of course not perceptually noticeable in speech, musical intervals and tonal hierarchy are subtly present in human daily vocalization. By analysing the amplitude-frequency combination of formants in human speech, Schwartz, Howe and Purves (2003) were able to predict both the structure of the chromatic scale and consonance/dissonance ordering – that is, they predicted the way people across several populations tend to provide similar judgment of a given interval in terms of consonance and dissonance. Schwartz et al. argue that this subconscious capacity would allow listeners to respond appropriately to significant sources of information embedded in human vocalisation. Such information would concern not only language itself but also cues such as the sex, age, and emotional state of the speaker.

Interval Name	Semiton Distance	Cents (12-TET)	Frequency Ratio
Unison (Un)	0	0	1:1
Minor 2nd (m2)	1	100	16:15
Major 2nd (M2)	2	200	9:8
Minor 3rd (m3)	3	300	6:5
Major 3rd (M3)	4	400	5:4
Perfect 4th (P4)	5	500	4:3
Tritone (TT)	6	600	45:32
Perfect 5th (P5)	7	700	3:2
Minor 6th (m6)	8	800	8:5
Major 6th (M6)	9	900	5:3
Minor 7th (m7)	10	1000	9:5
Major 7th (M7)	11	1100	15:8
Octave (Oct)	12	1200	2:1

Table 1. Musical intervals, expressed through their distance in semitones (s/t) from a given note, their sizes in cents for 12-tone equal temperament, and the corresponding mathematical ratios between two frequencies (in Hertz).

Focusing on fundamental frequency and formant relationships, Bowling et al. (2010) found that spectral analysis parallels the distinction between major and minor music tonal features. Because the physiological differences between excited (i.e., happy, bright) and subdued (i.e., sad, dark) affective states alter the spectral content of voiced speech (Scherer, 2003), the spectra of major intervals are more similar to

the spectra found in excited speech, while the spectra of particular minor intervals resemble the spectra of the subdued speech counterpart. In a second study concerning musical intervals in different languages and cultures, Bowling et al. (2012) found stronger similarities to music intervals in F0 than in spectral ratio analyses. It is worth mentioning that the studies so far presented in this section share a use of 'musical intervals' as a categorization matrix that allows any possible given ratio to be assigned to an interval.

A different case is that of Curtis and Barucha (2010), who found strong relationships between affective states and F0 ratios contained in speech contour: vocally generated intervals seemed to encode emotion, most noticeably in the case of sadness and the minor third. Here, 'music intervals' exclusively refer to ratios precise enough as to be considered as what is traditionally conveyed by the term in the musical domain. Curtis and Barucha's (2010) interpretation of these findings implicates causality: the arrangement of frequencies in music would, to some extent, mirror natural human vocalisation, placing language as a mapping source for the use of such parameters in music and hence underlying its affective impact. Along similar lines, Bowling et al. (2010) have suggested that the characteristic sentimental impressions of major and minor intervals would arise from routine associations

between specific musical intervals and voiced speech. What all of the previous studies have in common is the construction of ratios out of the pitch of subsequent syllable. However, intervals in tonal music not only entail representations of single events and local relationships on short time scales, but also nested hierarchical structures spanning longer time scales up to the entire length of a piece (Schenker, 1956; Salzer, 1962). Such non-local dependencies require working memory, do not require explicit structural knowledge, cannot be explained by Markov models (Koelsch, Rohrmeier, Torrecuso, and Jentschke, 2013) and remain to be explored in the conext of speech.

Approaching interpersonality

These studies on musical intervals in speech also share the methodological feature of approaching the phenomenon at a single person level of analysis, revealing a transmissional paradigm of communication (Kashkin, 2012), and ignoring the notions of feedback (Wiener, 1961) or dialogicity (Bakhtin, 1981) that the interactional paradigm implies (Kashkin, 2012). At the same time, musical intervals are not only generated within an instrument but also *among* instruments (allowing duet, ensemble, or choir performances). Considering these facts, a methodological

focus on a single person may be a pitfall when studying musical intervals in human speech. In a conversational context, it is indeed reasonable to expect that pitch relations will be found not only within a single speaker's F0 but also when juxtaposed to the F0 emitted by the speaker's interlocutor.

Following this line of reasoning, diverse prosodic phenomena have been reported in real dyadic interactions, particularly some that are focused on either fundamental frequency or F0 (Kramer, 1964; Johnson, 2003). By performing conversational and phonetic analysis on whole sequences of two-party conversations, Ogden (2006) qualitatively describes the way 'speakers match their own tone production to that of another speaker and manipulate the relation between their co-participant's production and their own in ways that have implications for meaning' (p.1773). The phenomenon known as phonetic convergence (Kim & Horton, 2011; Gregory, Dagan, and Webster, 1997) is also of special interest. It is usually defined as the progressive accommodation of the distance (in Hz) between a speaker's and his interlocutor's F0s. Recently, a close relationship between the level of empathy shared by two interacting parties and the corresponding level of phonetic convergence unfolding during interaction has been reported (XXX, 2012). In other words, a

psychological interactional disposition such as empathy influences the ratios between frequencies in speech prosody.

However, given that these studies do not accurately describe such ratios, it remains unknown whether musical intervals play a role in interactional situations. Hence, considering that the distance between two F0s can be modelled through mathematical ratios in the same form as the relation between two tones of a musical interval, given the evidence of musical intervals within a single person's speech, and the fact that nonlocal dependencies are a key component both in language[i] —at a syntactic level—(Nevins, A., Pesetsky, D., Rodrigues, C. (2009) and music (Koelsch et al., 2012), we hypothesised reciprocal nonlocal F0 accommodations from a musical perspective between persons during a conversational interaction. By manipulating psychological disposition, we expected to see differences among conditions in terms of interpersonally generated musical interval distribution and/or accuracy .

Methods

Participants

A total of 56 undergraduate students (30 female and 26 male) from the XXX voluntarily participated in the study. These individuals ranged in age from 18 to 28

years old, with a mean age of 21.04 years (SD 3.11). They were invited to take part in face-to-face conversations with an unknown participant as partner. Of this sample, 28 dyads of participants unknown to each other were randomly generated: ten female-female, seven male-male, and ten mixed-gender dyads. Each participant received a lunch voucher as compensation for participating.

Procedure

Prior to each session, experimenters separately scheduled two randomly paired participants, inviting them to join an everyday conversations study. Once they arrived, participants read and signed an informed consent form that had been previously reviewed and approved by the proper committees. Afterwards, both participants were taken to the XXX Laboratory, where individuals who had previously met one another were discarded to eliminate biased partners. No dyad declared previous knowledge of each other. The conversation room included two stools, placed 5 feet apart, on which participants could sit during the interaction. Wireless headset microphones were provided to each participant.

The total sample of dyads was divided into two experimental conditions, Low Trust and Control, which were balanced by number and sex –the latter controlled as it does not concerns this study. All dyads engaged in conversations guided by an

adaption of the Fast Friends Questionnaire (Aron, Melinat, Aron, Vallone, & Bator, 1997) to maintain a common structure among different dyads' conversations. Conversations consisted of a series of mutual questions about participants' names, families, and hobbies. Questions were translated and adapted to include culturally relevant questions (e.g., 'What happened to you during the February 2010 earthquake[ii]? How did you get through it?'). Once participants were seated, experimenters handed them a copy of the conversation instructions and read them aloud. Instructions were the same for both groups: 'Now you will talk about the questions written on these cards [two identical stacks of 10 printed cardboards with the questions]. If you wish, you can ask additional questions, but the aim is to focus on these questions written on the cards. The whole conversation should take from 20 to 40 minutes.'

In both conditions, the task consisted of getting to know each other through the proposed questions, and participants were informed that they would be asked about this information afterwards. In the Control Condition, participants were invited to 'just talk'. In the Low Trust Condition, they were additionally warned about the fact that their partner may have been instructed to lie on at least one of his answers. Depending on the type of interaction to which a particular dyad was previously

assigned, the subsequent instructions varied as follows. For the Control Condition (henceforth 'CC') group instructions stated: 'We ask you to answer these questions as spontaneously as you can and to listen carefully to your partner's answers. Try to get to know each other based on the answers.' Based on this kind invitation, participants were expected to feel confident and share naturally.

For the Low Trust Condition (henceforth 'LTC'), however, the instructions differed as follows: 'As you may know, in conversations people do not always tell the truth, so we tried to incorporate this fact in this study. Each one of these cards has printed on it a question but also a particular instruction for answering it: tell the truth, or lie. We don't know which card set you will choose, but one of them will have cards asking you to provide an answer that is not totally honest. Now, please take a set of cards without letting your partner know what your instructions are.' Without the participants' knowledge, both set of cards had the honesty instruction (i.e., 'Tell the truth in this answer'). Consequently, according to the instructions received, all participants in the LTC answered as truthfully as those in the CC, but believing that their conversational partner had lied to them: because they received the 'honest' set of cards, the partner's set necessarily had to be the 'non-honest' one. Trust has been traditionally defined as a series of evaluations and beliefs from which a person is likely to

be reliable, cooperative, or helpful (Simpson, 2007). Luhmann (1998) distinguished (reflexive) trust from confidence, which stands as 'pre-reflexive trust' in which people rarely intervene, and becomes visible only as it is interrupted. Accordingly, the introduced non-honest element implied one main effect. Acting as a transgression or betrayal event, believing that the partner had the 'non-honest' set was expected to break participants' confidence in one another, forcing them to start reflexively evaluating trust, or not trusting at all, thereby generating the previously mentioned restraint of prosodic expressions of emotion (Saarni, 1998) and more inhibited vocal expressions (Trainor et al., 2000).

This manipulation was expected to introduce a barrier in bonding for the LTC dyads. By forcing them to continuously think over and examine their partners' statements, they would be distanced from the here-and-now exchange, and they would inhibit their interactions, genuine emotional intentions, and empathic concern. When the conversation ended, participants received their vouchers. In the case of the LTC group, a debriefing took place, in which participants were informed of their partners' actual instructions.

Materials

Each participant's vocal emissions during the conversation were recorded on separate channels using individual SYSTEM 8 Audio-Technica headset microphones and digitalised through external sound card Focusrite Scarlett 8i6.

Speech samples consisted of whole sections from the entire conversations. As done in previous studies (XXX, 2012), participants' vocalisations from the beginning and ending sections were selected, corresponding to the conversations elicited by questions number 1, 2, and 9 from the Fast friends questionnaire. This allowed for the sampling of various moments, thus avoiding possible section biases. A total of 162 section vocalisations were analysed (3 per participant across 54 conversational dyads). Each section contained an average of 9.57 conversational turns per participant, each turn containing an average 43.08 vocal nuclei and surrounding FO contour.

Average vocal sample length (the sum of each participant's vocalisations corresponding to questions 1, 2, and 9) per condition was assessed, resulting in 1.28 minutes for the LTC and 1.82 for its counterpart. The durations tended to be smaller for the LTC: the effect of condition on log-duration was statistically significant in an ANOVA model at alpha=10% [($F(1, 52)=3.001$, $p=0.0891$)] when considering different

questions to be a repeated measure. Despite this slight difference, the amount of sound information per participant in both conditions by far exceeded most studies' sample length, which generally comprises a few words or short sentences (Curtis and Barucha, 2010; Bowling, Gill, Choi, Prinz, and Purves, 2010; Bowling, Sundararajan, Han, & Purves, 2012). It is worth noting that the audio data did not consist of monologues, shadowing or repeating texts, or artificially constructed sentences, but it instead consisted of a real conversation between two people.

Acoustic analysis

For investigation within a single person, there are several possible ways of contrasting discrete pitch measures of a melodic contour (Curtis and Barucha, 2010; Bowling et al., 2012). When approaching vocalisations between two persons, however, logical possibilities multiply, and theoretical ones are lacking. Although previous authors may have managed to construct criteria for pairing frequencies (and thus generating intervals) within a single person's speech, it is less evident how one should pair frequencies between two interacting speakers.

[Insert Table 2]

Table 2. Mode pairing and assimilation to musical interval categorization matrix. Grouped by colour (different shades of grey), pitch modes corresponding to a particular dyad's two participants (e.g., Control Condition dyad number 1, participants A and B) were paired in each question (q1, q2, and q9). Each pairing generated a mathematical ratio assimilated to a particular 12-TET interval. (e.g., Aq1 and Bq1's ratio is 1.505611972, corresponding to the interval of a Perfect Fifth or P5).

Using the PRAAT software (Boersma, 2001), each section's F0 mode was calculated in Hertz using a custom-made script. The analysis parameters were based on the F0 range of adult speech, corresponding to an F0 calculation range of 90 to 800 Hz in the case of women and 65 to 600 for men. The F0 contour of each speech sample was calculated using the Praat autocorrelation method, with a voicing threshold of 0.45 and a frame period of 0.005 seconds. Mode was selected as an important and representative parameter because it consists of the most repeated and stable frequency and is stronger when considering vast amounts of values, as is the case for a whole section of utterances. Mode was also selected because, unlike the mean, it displays a real F0 frequency value that participants actually emitted.

Because the F0 mode corresponded to each dyad participant's series of utterances during a particular question, it was possible to pair them (e.g., Participant A's F0 mode corresponding to question 9 utterances paired with Participant B's F0

mode corresponding to question 9 utterances). Out of each F0 mode pair, a mathematical ratio was calculated in cents[iii]. Because musical intervals are perceived in semitones in the occidental cultural background, with each semitone comprising 100 cents, such mathematical ratios were then categorised into 100-cent bins. Intervals larger than an octave (1 in the CC, 5 in the LTC) were normalized to within octave range for all categorical analyses, as customary in the conceptualization of pitch class hierarchies in tonal music. The literature identifies perceptual discrimination boundaries at 50 cents between musical interval category prototypes (Burns & Campbell, 1994; Burns & Ward, 1978; Siegel & Siegel, 1977). Because such boundaries were recently corroborated in the specific field of vocal musical intervals (Curtis et al., 2010), 50-cent bin categorisation was also performed. Therefore, the main pitch distance between interacting partner's vocalisations could be analysed through musical logic.

Results

Using 100-cent binning as a categorization matrix, differences were found among conditions. As shown in Figure 1, participants tended to vocally assimilate to one another, generating different mathematical ratios. In the case of the LTC a large

percentage of pitch mode ratios corresponded to small intervals, in particular, minor and major seconds. Few larger intervals were found in this group. In the CC, small intervals were also present, but these appeared along with larger intervals, such as sixths and the minor seventh. Therefore, a larger and more even distribution was observed in the CC. When clustered into two groups, namely smaller than a tritone and equal or greater than a tritone, a χ_2 analysis confirmed such distribution differences [χ_2 (1) = 3.89, p = 0.048].

Given such differences, it seemed reasonable to ask whether such differences may have arisen from participants' individual vocal characteristics. CC dyads may have comprised deep-voiced men talking to shrill-voiced women, thus generating larger intervals, or LTC dyads comprising shrill voiced men talking to deep voiced women may have generated smaller intervals. Consequently, the pitch range was calculated for participants in both conditions. These values were compared using Student's t tests for independent samples, assuming equal variance. There was no evidence of a significant between-group difference in tonal range [t (22) = 0.5583, p = 0.5822]. This result indicated that individual's physiological/vocal characteristics did not explain these findings.

[Insert Figure 1]

Figure 1. Differences among conditions on interval distributions using a 100 binning (smaller, or equal/larger than a Tritone) were statistically significant [χ_2 (1) = 3.8866, p = 0.04867]. LTC: Low-Trust-Condition; CC: Control Condition.

[Insert Figure 2(a)]

[Insert Figure 2(b)]

Figure 2. Differences among conditions on interval distributions using a 50 binning (smaller, or equal/larger than a Tritone) were statistically significant**. (a)** comprises ratios within the 50-cent binning, hence labelled as 'tuned', [χ_2 (1)=26.337, p<0.001]. **(b)** comprises ratios outside the 50-cent binning, labelled as 'un-tuned' [χ_2 (1)=3.689, p=0.055]. LTC: Low-Trust-Condition; CC: Control Condition.

By utilising stricter categorisation for mode ratios such as 50-cent binning, the results can be better understood. As presented in Figure 2 (a and b sections), for mode ratios that did not satisfy such binning and have thus qualified as perceptually ambiguous in the literature, very slight differences among conditions can be observed

(Fig. 2b). Therefore, these ratios are henceforth tentatively called 'un-tuned'. Conversely, when considering only those ratios that can actually be labelled as 'musical intervals', or 'tuned' (Fig. 2a), the condition dissimilarities are enhanced. The distribution of 'tuned' and 'un-tuned' evidently showed no significant difference among conditions (Figure 3)

[Insert Figure 3]

Figure 3. Percentage of intervals that satisfied ('tuned') and did not satisfy ('un-tuned') the perceptual discrimination boundary at 50 cents for the musical interval category. No significant differences between conditions were found. LTC: Low-Trust-Condition; CC: Control Condition.

Pearson's χ^2 tests with Yates' continuity correction were performed separately for tuned and un-tuned intervals to establish whether the proportion of smaller vs. equal or greater than tritone intervals was different for the two experimental conditions (CC and LTC). In fact, a very strong between-condition difference was found in tuned interval data [χ^2=26.337, df=1, p<0.001], whereas a less strong difference was found in the un-tuned interval data [χ^2=3.689, df=1, p=0.055]. To express meaningful effect sizes, odds for the larger than tritone intervals were computed. Doing so yielded an odds ratio of 5.14 for tuned intervals and 1.85 for un-

tuned. In other words, if we look only at tuned intervals, the odds of larger than tritone intervals are more than five times higher in CC than in LTC. When looking at tuned data, the odds are still higher for CC but are less than twice of those for LTC. As a result, the pattern of more frequent large intervals in CC than in LTC was statistically significant and very clear within tuned intervals data, but not to the same extent within 'un-tuned' intervals.

By restricting generated musical intervals exclusively to those within a 50-cent binning, thus considering ratios susceptible of being perceptually discriminated as recognisable or tuned, some similarities were found when comparing the results to those of Figure 1. Nevertheless, there were also interesting differences. First, among the CC's tuned intervals, none qualified as a minor or major second (see Figure 2(a)). This strongly contrasts with LTC's clear preponderance of such intervals. Second, the CC's tuned intervals showed a marked and sudden peak of minor thirds and then remained on a lower plateau until the major sixth, where it picked up again. The opposite behaviour was observed in the LTC. The average pitch distance among the interlocutors' F0 modes, per section, showed no significant difference between LTC's first (1.41) and last (1.44) sections. On the other hand, there was a significant

reduction of the average pitch ratio in the CC (1.61 and 1.22) (t = 12.9498, df = 10, p-value = 1.423e-07).

Discussion

Our results demonstrate a relationship between psychological disposition and the distribution of musical intervals between persons participating in a real conversation. We found that tones corresponding to interlocutors' F0s spontaneously generate ratios that can be categorised as musical intervals. This finding does not necessarily imply that mode frequency ratios in speech are *perceived* as musical intervals, but it does strongly suggest that conversational frequency patterns are related to Western music intervals. Such information highlights novel relationships between psychological disposition and vocalisation because the interval distribution differed among conditions.

The main finding of this work concerns categorical accuracy or 'tuning', which stands as a key element of the differences found between conditions. The interval distributions of the LTC displayed in Figures 1 and 2 reveal a similar pattern: a greater presence of smaller intervals. This predominance has also been observed by Bowling et al. (2012), who compared intervals—regardless of their ratio accuracy—smaller

and larger than a major second, which was chosen as a cutoff because it maximised the differences between positive/excited and negative/subdued emissions. It could be hypothesised that smaller intervals are a predominant feature of regular vocal emissions, both in melodic contour and interpersonally juxtaposed pitches.

However the same cannot be said for the CC. When contrasting the CC's 50-cent binning (tuned) interval distribution (Figure 3) with the 100-cent and 50-cent 'un-tuned' distributions (Figures 1 and 2), the patterns differ importantly. Whereas major and minor seconds were present in the 100-cent binning, as were intervals larger than a tritone (more similar to the LTC in all of its versions), there was a remarkable absence of seconds in the 50-cent binning. Instead, a preponderance of larger intervals was not only maintained but also enhanced. Furthermore, the un-tuned version of the 50-cent distribution shows a regularity that diverges very little from all the binning versions of the LTC. It is worth mentioning that without the octave normalization, the raw size of the intervals in the CC shows smaller intervals in general when compared to the LTC. Also, the significant pitch ratio reduction along CC conversations implies phonetic convergence, as would have been expected.

These findings demonstrate that psychological dispositions modulate musical intervals between conversational partners. As the LTC's manipulation broke

interpersonal confidence, it inhibited interactants' expressions, inducing a reflective disposition toward the other person (Trainor et al., 2000). This would not have occurred in the CC, where less constrained prosody could arise from both speakers, thereby generating the unique 50-cent tuned distribution.

Because previous research from Bowling et al. (2012) and Curtis and Barucha (2010) had an exclusively single-person scope, it is rather improbable that these results can be explained by single-person patterns that are reflected merely by chance at the interpersonal level. Thus, though the intervals studied here were generated by comparing two different persons' pitches, the possibility that the predominance of small intervals (primarily seconds) could still be explained as a consequence of physiological phenomena is improbable, although this possibility cannot be discarded and should therefore be explored. However, our findings instead suggest a cognitive connection between the individual physiological characteristics of human sounds and the interpersonal psychophysiological level. In particular, our results stand as evidence of how the motivational-structural dimension of sound might afford its socio-intentional role (Cross, & Woodruff, 2008). According to these authors, music's socio-intentional dimension is oriented towards the interpretation of human agency and intentionality. It arises from sound structures that could be interpreted as

affording cues about shared intentionality that direct attention in interaction. Such sound structures may be construed as disclosural or dissimulative, denoting distinct and different communicative intentions and hence establishing the pragmatic contexts of utterances. Although originally conceived in engagement with music, a socio-intentional dimension can also be considered in vocal pitch interplay. In this particular case, the same sound structures (pitch ratio patterns) contained in melodic contour as an involuntary physiological 'honest signal' can be found within two persons' vocalisations. However, at interactional level, they are no longer exclusively physiological but are socially constructed structures. Only mutual vocal accommodations that require complex online cognitive processing allow interpersonally generated musical intervals to happen. The fact that the socio-intentional dimension of vocal sound conveys a disclosural or dissimulative character is also important. Because the conditions differed through the introduction of lies, it is sensible to hypothesise that interpersonal pitch distances played a role in that matter and thus also differed among conditions.

When considering interpersonal pitch phenomena, another topic of discussion is how it is possible to coordinate such relations. Among Hockett's (1960) design features of animal communication, two features are particularly interesting in this

context. The first is Interchangeability, by which a speaker can both transmit and receive at the same time. The second is Total Feedback, which means that the speaker hears what he himself says. Hockett highlights the importance of the latter because it allows for the internalisation of communicative behaviour, which constitutes a major portion of what we call 'thinking'. Therefore, it is possible that by obtaining a certain notion of the position of one's own sounds within the larger sound plot of an interaction, it would be possible to insert one's own vocal/oral/communicative activity within interpersonal social dynamics. Also highlighting our capacity to vocalise within acoustic contexts, Bannan (2008) claims that '[H]arnessed to a 'theory of mind', a theory of harmony acts as a bridge between musical vocalisation and the properties of categorical discrimination that are demanded by unambiguous language…The theory of mind that permits unison- and harmony-singing comprises a unique cognitive adaptation that could be seen as the candidate human equivalent to the capacity of birds instinctively to fly in flocks and fish to shoal for their collective protection' (Bannan, 2008, p. 285). Although thinking of choirs rather than interpersonally generated intervals, Bannan stresses the psychological and social implications of the human ability for perceiving not only the tones we emit but also their relation with other peoples' tones – namely, musical intervals in the manner of a

dynamical system (Kelso, 1995). Bannan's notion of *complementary pitches* becomes crucial because it implicates that both interactants' own tonal behaviour as 'speakers' is not lost, but rather registered along with the one they listen to as 'hearers'. This idea is congruent with Ogden's (2006) statements. After his sequential analysis of agreement and disagreement conversations, he concludes that meaning conveyed through phonetic features is a local and collaborative achievement. The implication of this claim is that some linguistic features should be considered in their local context – in his and in our case, a dialogical one. This type of 'situated' (Malloch & Trevarthen, 2009) or 'attitudinal' (Cruttenden, 1997) meaning would then not reside solely in one of the interactants' minds, as 'trust' and 'mistrust' only have meaning in relation to at least two persons.

One additional consideration with regard to the present findings is pitch discrimination accuracy. As stated above, our results neither imply nor deny that mode frequency ratios in speech are consciously *perceived* as musical intervals. In fact, it seems difficult to imagine that interactants would be able to conciously relate each other's vocal sounds to whole supra-segmental sections, as in this case. Still, the differences among conditions are consistent and confirm the idea that nested nonlocal dependencies is a multi domain capacity of human cognition that relies in working memory

(Koelsch et al., 2012). As it has already been demonstrated to be present both in tonality and speech syntax, it is not surprising to also find it in prosody (F0). Zatorre & Baum (2008) affirmed that despite their shared cognitive processes, the way in which pitch information is handled differs between speech and music. They propose two mechanisms: one that is 'coarse-grained', focused on contour, and another that is 'fine-grained', involving more accurate pitch relations such as musical intervals. The first one might overlap across domains; the second is hypothesised to be specifically music-directed. The present findings provide relevant contributions to examine such propositions. Differences among the LTC and CC were only conclusive when considering 50-cent 'fine-grained' pitch ratios. The presence of this mechanism – thought to be exclusively musical – becomes a necessary element for the unfolding of the adjacent pitch dynamics reported in this study. Because Zatorre & Baum's evidence considers only melody and sentence-level intonation, their differing hypothesis might be accurate at such a level, but the present evidence demonstrates otherwise at the supra-segmental level. This finding highlights the importance of paying attention not only to ruled-based, grammatical sub-segmental aspects of language but also to its holistic aspects (Wray & Grace, 2007). In particular, it is important to consider that pitch dynamics (both on a single or dyadic approach) may

concern not only what can be noticed through melodic contour or adjacent pairs but also that which is beyond the structure that grammatical rules suggest, including supra-segmental data.

Future research should focus both on exploring such holistic aspects – for example, where on a finer, microgenetic level do such pitch ratios occur during interactions. As the socio-intentional dimension of music arises not only from sound but from several performative actions, it would also be interesting to investigate what other pragmatic aspects of language (gesture, body movement) may be related to the vocal phenomena presented in this work, and how they would link to each other.

Limitations

The mode extraction method utilised in the experiments only guarantees the consideration of actually present tones, which in some cases may correspond not to the F0 but to its formants. This opens the question of what the representative note of these segments should be, which can be explored in future work.

Ethical Approval

Ethical approval for this project was given by the Ethical Committee of the Psychology Department of XXX as well as by the Bioethics Advisory Committee of XXX [Grant 1100863].

References

Aron, A., Melinat, E., Aron, E.N., Vaollone, R., & Bator, R. (1997). The experimental generation of interpersonal closeness: a procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, *23*, 363-377.

Bakhtin, M.M. (1981) *The Dialogic Imagination: Four Essays*. Ed. Michael Holquist. Trans. Caryl Emerson and Michael Holquist. Austin and London: University of Texas Press.

Bannan, N. (2008). Language out of music: the four dimensions of vocal learning. *The Australian Journal of Anthropology, 19*(3), 272-293.

Bidelman, G., & Krishnan, A. (2009). Neural correlates of consonance, dissonance, and and the hierarchy of musical pitch in the human brainstem. *The Journal of Neuroscience, 29*, 13165–13171.

Boersma, P. (2001). Praat: a system for doing phonetics by computer. *Glot International, 5*, 341–345. Retrieved from http://www.praat.org

Bowling, D. L., Gill, K., Choi, J. D., Prinz, J., & Purves, D. (2010). Major and minor music compared to excited and subdued speech. *Journal of the Acoustic Society of America*, *127*(1), 491–503.

Bowling, D.L., Sundararajan, J., Han S., & Purves D. (2012). Expression of Emotion in Eastern and Western Music Mirrors Vocalization. *PLoS ONE, 7*(3), e31942. doi:10.1371/journal.pone.0031942

Brown, S. (2000). The "musilanguage" model of musical evolution. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The origins of music* (pp. 271–300). Cambridge, MA: The MIT Press.

Cross, I. & Woodruff, G. E. (2008). Music as a communicative medium. In Botha, R. & Knight C. (Eds.), *The prehistory of language* (pp.77-98). Oxford: Oxford University Press.

Cruttenden, A. (1997). *Intonation*. Cambridge, UK: Cambridge University Press.

Curtis, M., & Barucha, J. (2010). The minor third communicates sadness in speech: mirroring its use in music. *Emotion*, *10*(3), 335–348.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Development, 60*, 1497–1510.

Femald, A. 1992. Meaningful melodies in mothers' speech to infants. In H. Papousek, U. Jürgens, & M. Papousek (Eds), *Nonverbal vocal communication: comparative and developmental approaches* (pp.262-282). Paris: Éditions de la Maison des Sciences de l'Homme.

Fernald, A. (1991). Prosody in speech to children: prelinguistic and linguistic functions. *Annals of Child Development, 8*, 43–80.

Gregory, S.W., Dagan, K., & Webster, S. (1997). Evaluating the relations between vocal acommodation in conversational partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, *21*, 23-43.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 89-96.

Johnson, K. (2003). *Acoustic and auditory phonetics*, 2nd ed. Oxford: Blackwell.

Kashkin, V. B. (2012). Telementation vs. Interaction: Which Model Suits Human Communication Best?, *Journal of Siberian Federal University*. Humanities & Social Sciences, 12(5), 1733-1743.

Kelso, J. S. (1995). Dynamic Patterns: The Self-Organization of Brain and Behavior. The MIT Press.

Kim, M., & Horton, W.S. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, *2*, 125-156.

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.

Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal and Social Psychology*, *68*, 390-396.

Luhmann, N. (1998). Familiarity, confidence, trust: problems and alternatives. In D. Gambetta (ed.), *Trust: making and breaking co-operative relations*, pp. 94–107. Oxford: Blackwell.

Malloch, S., & Trevarthen, C. (2009b). Musicality: communicating the vitality and interests of life. In S. Malloch, & C. Trevarthen (Eds.), *Communicative musicality: exploring the basis of human companionship* (pp. 1–11). Oxford: Oxford University Press.

Masataka, N. (2006). Preference for consonance over dissonance by hearing newborns of deaf parents and of hearing parents. *Developmental Science*, *9*(1), 46-50.

Nevins, A., Pesetsky, D., Rodrigues, C. (2009). Pirahã exceptionality: A reassessment. *Language*, 85, 355–404.

Ochs, E. (1996). Linguistic resources for socializing humanity. In J.J. Gumperz, & S.C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 407-437). Cambridge: Cambridge University Press.

Oelmann, H., & Laeng, B. (2008). The emotional meaning of harmonic intervals. *Cognitive Processing, 10,* 113-131.

Ogden, R. (2006). Phonetics and social action in agreements and disagreements. *Journal of Pragmatics*, *38*, 1752–1775.

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). doi: 10.3389/fnhum.2012.00026

XXX, (2012). XXX. *XXX*, *50*(2), 145-65.

Salzer, F. (1962) Structural Hearing: Tonal Coherence in Music (NY: Dover), Vol 1.

Spreng, R.N., McKinnon, M., Mar, R., & Levine, B. (2009). The Toronto Empathy Questionnaire: scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personal Assessment*, *91*(1), 62-71.

Schellenberg, E., & Trehub, S. (1999). Natural musical intervals: evidence from infant listeners. *Psychological Science, 7*, 272-277.

Schenker, H. (1956) Neue Musikalische Theorien und Phantasien: Der Freie Satz (Vienna: Universal Edition), 2nd Ed.

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication*, *40*, 227–256.

Schwartz, D. A., Howe, C. Q., & Purves, D. (2003). The statistical structure of human speech sounds predicts musical universals. *The Journal of Neuroscience*, *23*(18), 7160 –7168.

Simpson, J. (2007). Psychological foundations of trust. *Current Directions in Psychological Science*, 16, 264-268.

Smith, L., & Williams, R. (1999). Children's Artistic Responses to Musical Intervals. *The American Journal of Psychology*, *112*, 383-410.

Stern, D.N., Spieker, S., & MacKain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Developmental Psychology, 18*, 727-735.

Trainor, L., Austin, M., & Desjardins, R., (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, *11*(3), 188-195

Trainor, L., & Desjardins, R. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*(2), 335-340.

Trainor, L., Christine D. Tsang, C., & Cheung, V. (2002). Preference for sensory consonance in 2- and 4-month-old infants. *Music Perception*, *20*(2), 187-194.

Trehub, S.E., Trainor, L.J., & Unyk, A.M. (1993). Music and speech processing in the first year of life. In H.W. Reese (Ed.), *Advances in child development and behavior* (pp. 1-35). New York: Academic Press.

Trehub, S. (2003). The developmental origins of musicality. *Nature Neuroscience, 6*, 669-673.

Vijay, I. (2002). Embodied mind, situated cognition, and expressive microtiming in African-American music. *Music Perception: An Interdisciplinary Journal*, *19*(3), 387-414.

Zatorre, R. J., & Baum, S. R. (2012). Musical melody and speech intonation: singing a different tune? PLoS Biology, *10*(7), e1001372. doi:10.1371/journal.pbio.1001372.

Wiener, N. (1961). *Cybernetics, or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.

Wray, A. & Grace, G. W. (2007). The consequences of talking to strangers: evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*, 543-578.

Footnotes

[i]

[ii] On February 27, 2010 an earthquake occured off the coast of central Chile with a magnitude of 8.8 Richter grades.

[iii] Following the Cents =1,200 [log(f1/f2)/log2] formula.