Redundancy makes music and speech robust to individual differences in perception.

.

Kyle Jasmin[1-3], Fred Dick[1,4], and Adam Taylor Tierney[1]

1. Department of Psychological Sciences
Birkbeck, University of London

2. UCL Institute of Cognitive Neuroscience

3. Laboratory of Brain and Cognition
National Institute of Mental Health, NIH

4. UCL-Birkbeck Centre for Neuroimaging

Correspondence to:
Kyle Jasmin
Department of Psychological Sciences
Birkbeck, University of London
Malet Street, London
WC1E 7HX
k.jasmin@bbk.ac.uk

Abstract

Communicative auditory signals express their structure through acoustic dimensions such as pitch and timing. Individuals' abilities to perceive these dimensions vary widely, and yet most people seem to comprehend music and speech easily. How? Here we tested whether redundancy – multiple acoustic cues indexing the same feature – makes music and speech robust to such individual differences. A model population with a severe and specific deficit for perceiving pitch (congenital amusics) and controls completed 3 tasks, each testing whether they could take advantage of redundancy. In each task, performance relied on either pitch or duration perception alone, or both together redundantly. Results showed that when redundant cues are present, even people with a severe deficit for one type of cue can rely on another to improve their performance. This suggests that redundancy may be a design feature of music and language, one that assures transmission between people with diverse perceptual abilities.

Introduction

Auditory communication systems like music and language convey information through relatively continuous sound streams. But at an abstract level, these streams consist of smaller units (notes, motifs, words) combined hierarchically into larger structures (lines, phrases, sentences) (Lerdahl & Jackendoff, 1985). Comprehending structural aspects of these signals requires identifying how adjacent elements (like words in language, notes in music) are grouped, and how they relate to one another. This structural information is conveyed through variations in acoustic dimensions such as pitch and timing - but individuals differ substantially in their ability to perceive these (Grondin, 1993; Deguchi et al., 2012; Phillips-Silver, 2011). How then are communicative auditory signals like music and speech perceived so successfully, despite large individual differences?

The answer may lie in redundancy. Pitch, duration and amplitude changes often co-occur in time and provide cues to the same (rather than different) structural features. For instance, the boundaries of *musical phrases*, the smallest group of related adjacent units in music, are characterized by changes in pitch (a shift from low to high or high to low) and timing (a tendency for longer notes near a phrase's end; Palmer and Krumhansl, 1987). In language, *linguistic phrase* boundaries are similarly marked by a pitch shift from low to high, or high to low, and also by lengthened syllable durations (Figure 1C-D)(Streeter, 1978; Wightman et al., 1992). *Linguistic focus* (emphasis on a word) is also indicated acoustically by a pitch

excursion, durational lengthening and an amplitude increase (Figure 1A-B, Sluijter et al., 1996).

It has been suggested that this redundancy (or "degeneracy") of acoustic channels may be a design feature of speech that makes it robust to background noise in the environment (Winter, 2014; Patel 2014). Does redundancy also make speech robust to large individual differences in perceptual ability? And does redundancy also make *music* robust to such individual differences?

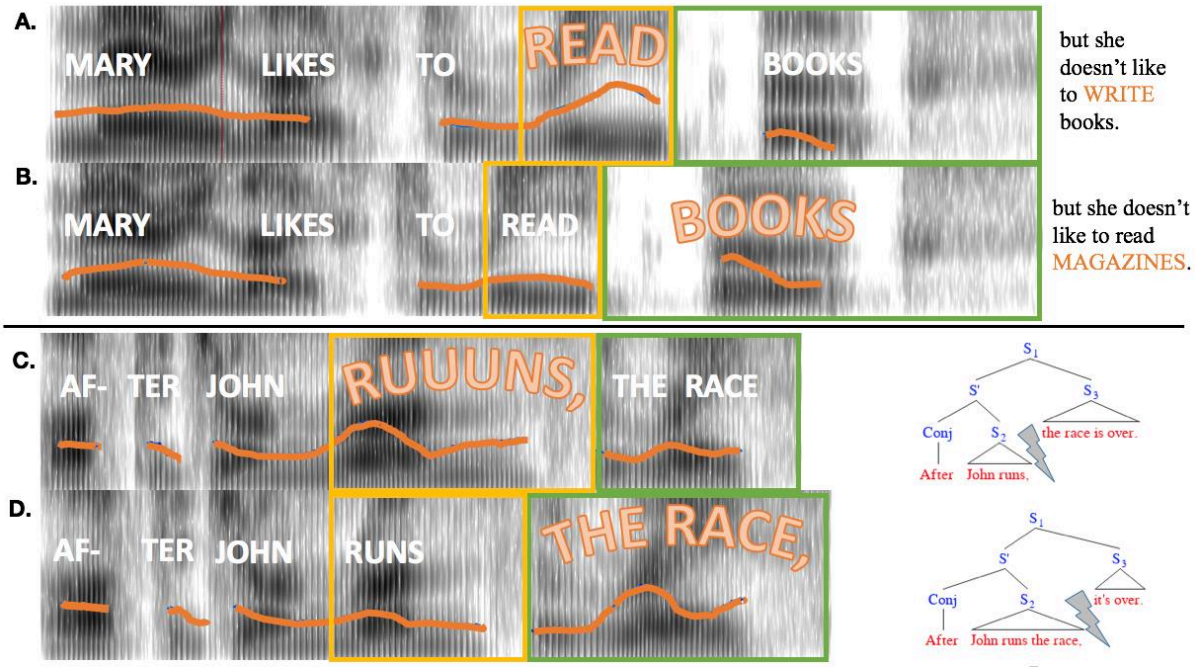

Figure 1. Pitch and duration correlates of emphatic accents and phrase boundaries. Spectrograms of stimuli used in the experiment, with linguistic features cued simultaneously by pitch and duration (the "Both" condition). Orange line indicates pitch contour. Width of yellow and green boxes indicate duration of the words within the box. A) emphatic accent places focus on "read". Completion of the sentence appears to the right. B) emphatic accent places focus on "books".

Sentence completion at right. C) a phrase boundary occurs after “runs”. D) a phrase boundary occurs after “race”. Syntactic trees are indicated at right, with phrase boundary indicated by jagged line.

Here we address these questions by examining perception of music and speech in a model population with a highly specific and extreme perceptual deficit. Congenital amusia is a non-clinical condition that is characterized by impaired processing of small changes in pitch and affects 1.5% of the population (Peretz & Vuvan 2017). Laboratory tests have shown that amusics have difficulty with distinguishing musical melodies based on pitch alone (Peretz, 2003). Amusics also sometimes struggle on pitch-related speech tasks (Patel et al., 2008, Hutchins et al., 2010; Nan et al., 2010, Jiang et al., 2010, Jiang et al., 2012; see Vuvan et al., 2015 for a meta-analysis), but not always (Ayotte et al., 2002; Peretz et al., 2002; Patel et al., 2005). In real-life situations, amusics may be able to compensate for their impaired pitch perception by relying on redundant cues to musical and prosodic features. If our model populations (with severe deficits) can take advantage of redundant cues in perceive speech, this would suggest that individuals with less severe deficits may also.

In an experiment on music perception, we examined whether amusics were able to make judgments about musical phrases when they could rely on pitch, duration, or both types of cues simultaneously. If amusics are able to take advantage of their unimpaired perceptual processing, their performance should be improved when they can rely on redundant cues (pitch and duration), compared to when they must rely solely on an impaired cue (pitch). Next, in two linguistic experiments, we measured the extent to which subjects used pitch and duration cues to perceive emphatic

accents ('Mary likes to *read* books, but not *write* them') and phrase boundaries ('After John *runs [phrase boundary]*, the race is over.'). To do this, we manipulated stimuli via voice morphing such that participants needed to rely on pitch cues alone, duration cues alone, or could use both together to perform the tasks. Given amusics' lack of self-reported language issues, we predicted that amusics would perform similarly to controls when they could also take advantage of duration cues (as in natural speech) but poorly on trials when they had to rely on pitch cues alone. Their ability to perceive speech in the presence of background noise was also assessed, which is known to be impaired in tone-language speaking amusics (Liu et al., 2015) but has not been established for speakers of non-tonal languages, who may rely more on duration cues when perceiving speech in noise (Lu & Cook, 2009).

Methods

Participants

Participants, 16 amusics (10 F, age = 60.2 +- 9.4) and 15 controls (10 F, age = 61.3 +- 10.4), were recruited from the UK and were native British English speakers with the exception of one amusic whose native language was Finnish but acquired English at age 10. This subject was excluded from the Linguistic Phrase and Focus Test analyses. No participant in either group had extensive musical experience. All participants gave informed consent and ethical approval was obtained from the ethics committee for the Department of Psychological Sciences, Birkbeck, University of London. Amusia status was obtained using the Montreal Battery for the Evaluation of Amusia (MBEA). Participants with a composite score (summing the Scale, Contour and Interval tests scores) of 65 or less were classified as amusics (Peretz et al., 2003). Amusia is a rare condition with 1.5% prevalence (Peretz & Vuvan, 2017).

The sample size was therefore limited by our ability to recruit, screen and test qualifying participants.

Musical Phrase Perception Test

The musical phrase perception test was designed to test our participants' ability to perceive how well a series of notes resembles a complete musical phrase.

Stimuli

The stimuli consisted of 100 musical phrases taken from a corpus of folk songs (Schaffrath 1995). They appeared in three conditions: Both – an unmodified version of the musical phrase; Pitch – where the pitch of the notes was preserved (as in the original version) but the durations were set to be identical, i.e. isochronous; and Time – where the original note durations were preserved but the pitch of the notes was made to be monotone. In an additional manipulation, half of the stimuli formed a complete musical phrase with the notes in an unmodified sequential order - these could be perceived as a *Complete* musical phrase. The other half were made to sound *Incomplete* by presenting a concatenation of the second half of the musical phase and the first half of the next musical phrase in the song. The order of the notes within the two halves was preserved. Thus the resulting *"Incomplete"* stimuli contained a musical phrase boundary that occurred in the middle of the sequence rather than at the end.

Procedure

On each trial, a stimulus note sequence was presented to the participant through headphones. After the sound finished playing, a response bar appeared on the screen which was approximately 10 cm in width. Subjects were tasked with deciding

how complete each musical phrase sounded by clicking with their mouse on the response bar. The word “Incomplete” was shown on the left side of the response bar, and the word “Complete” was shown on the right, and participants could click anywhere within the bar to indicate how complete they thought the phrase had sounded (Figure 2). After the participant indicated their response, the experiment continued, with the next stimulus being played immediately. Participants judged 3 blocks of 50 trials each with a short break in between. As the study was aimed at understanding individual differences, the block order was always the same, with all the trials in a condition presented in a single block (Both Cues, then Duration Only, then Pitch Only).

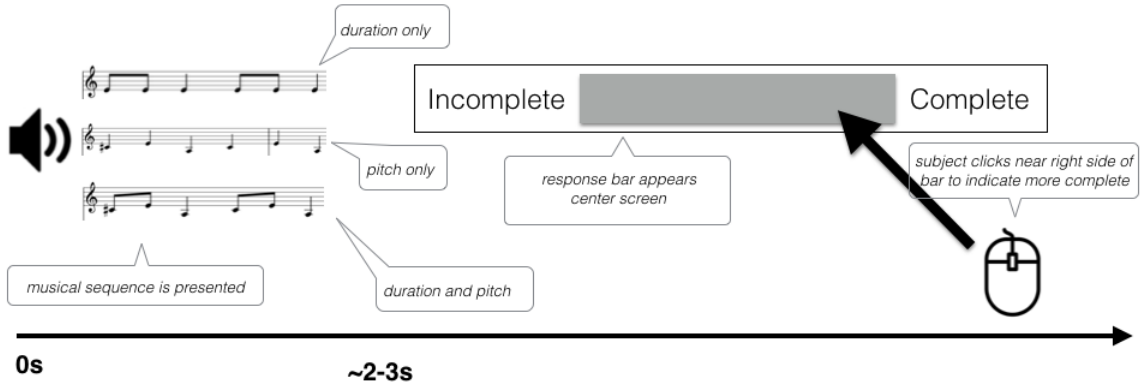

Figure 2. Schematic of trial structure for the Musical Phrase Test. Participants heard a musical sequence which was either a complete musical phrase or straddled a boundary of 2 musical phrases. They then indicated how complete they thought the phrase sounded by clicking with a mouse on a response bar.

Linguistic focus perception task

This test measured participants’ ability to detect where a contrastive pitch accent was placed in a sentence, based on only one auditory cue (Pitch or Duration) or both together (as in natural speech).

Stimuli

The stimuli consisted of 47 compound sentences with an intervening conjunction, e.g. “Mary likes to READ books, but she doesn’t like to WRITE books.” These were all created specifically for this study. Each of the sentences had two versions: “early focus”, where a word in all capital letters for emphasis (e.g. “READ”) occurred early in the sentence and served to contrast with a similar word later in the sentence, and “late focus”, where a similarly capitalized word occurred slightly later in the sentence (“Mary likes to read BOOKS, but she doesn’t like to read MAGAZINES”). Both versions of the sentence were lexically identical from the start of the sentence up to and including the conjunction (see Fig 1A,B).

We then recorded sentences as they were spoken by an actor, who placed contrastive pitch accents to emphasize the capitalized words. Recordings of both versions of the sentence were obtained, cropped to the identical portions (underlined above). Using STRAIGHT software (Kawahara & Tirino, 2005), the two versions were manually time aligned. We then produced 6 different kinds of morphs by varying the amount of pitch-related (F0) and temporal information either independently or simultaneously. For *pitch only* stimuli pairs, the late and early focus sentences differed only in pitch. The temporal morphing proportion between the two versions was held at 50% while the pitch was set to include 75% of the early focus version or 75% of the late focus version recording. This resulted in two new ‘recordings’ that differed in F0, but were otherwise identical in terms of duration, amplitude and spectral quality. For *duration only* stimuli, we created two more morphs that held the pitch morphing proportion at 50% while the temporal proportion was set to either 75% early focus or 75% late focus. The output files differed only in

duration, and but were identical in terms of pitch, amplitude and spectral quality. Finally, we made *"naturalistic"* stimuli where both pitch and temporal information contained 75% of one morph or the other, and thus pitch and duration simultaneously cued either an early or late focus reading.

Procedure

Stimuli were presented with Psychtoolbox in Matlab. Participants saw sentences presented visually on the screen one at a time, which were either early or late focus (see paradigm schematic in Fig 1 A,B and Fig 3A). The emphasized words appeared in all upper-case letters as in the examples above. Subjects had 4 seconds to read the sentence to themselves silently and imagine how it should sound if someone spoke it aloud. Following this, subjects heard the first part of the sentence spoken aloud in two different ways, one that cued an early focus reading and another that cued late focus. Participants were instructed to listen and decide which of the two readings contained emphasis placed on the same word as in the text sentence. After the recordings finished, subjects responded by pressing "1" or "2" on the keyboard to indicate if they thought the first version or second version was spoken in a way that better matched the on-screen version of the sentence. The correct choice was cued either by pitch or duration exclusively, or both together. The serial order of the sound file presentation was randomized. The stimuli were divided into 3 lists counterbalanced for condition and early vs. late focus.

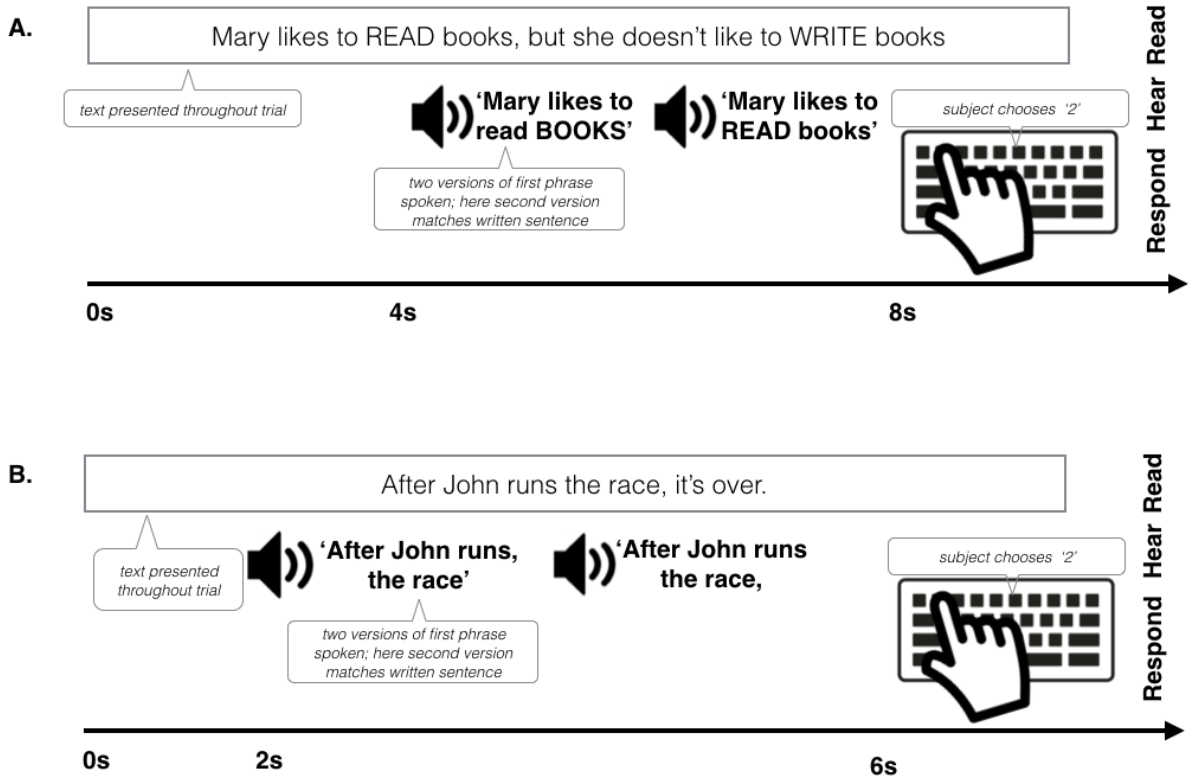

Figure 3: Example trial structure for the linguistic focus test (A) and the linguistic phrase test (B). First, a single sentence was presented visually, and the participants read it to themselves. Next, two auditory versions of the first part of the sentence were played sequentially, only one of which matched the focus pattern of the visually presented sentence. Participants then indicated which auditory version matched the onscreen version with a button press.

Linguistic phrase perception test

The Linguistic Phrase Perception Test measured participants' ability to detect phrase boundaries in speech which are cued by pitch only, duration only, or both pitch and duration.

Stimuli

The stimuli consisted of 42 short sentences with a subordinate clause appearing before a main clause. About half of these came from a published study

(Kjelgaard 1999) and the rest were created for this test. The sentences appeared in two conditions: an “early closure” condition, where the subordinate clause’s verb was used intransitively, and the following noun was the subject of a new clause; and “late closure”, where the verb was transitive and took the following noun as its object, causing the phrase boundary to occur slightly later in the sentence. Both versions of the sentence were lexically identical from the start of the sentence until the end of the second noun.

A native British English speaking male (trained as an actor) recorded early and late closure versions of the sentences. The recordings were cropped such that only the lexically identical portions of the two versions remained, and silent pauses after phrase breaks were excised. The same morphing proportions were used as before – with early or late closure cued by 75% morphs biased with pitch, duration or both. As before, the stimuli were crossed with condition and early vs. late closure and divided into three lists.

Procedure

The procedure for the Linguistic Phrase test was similar to that of the Linguistic Focus Test. Participants saw sentences presented visually on the screen one at a time, which were either early or late closure, as indicated by the grammar of the sentence and a comma placed after the first clause (Figure 2B). They then had two seconds to read the sentence to themselves silently and imagine how it should sound if someone spoke it aloud. Following this period, subjects heard the first part of the sentence (which was identical in the early and late closure versions) spoken aloud, in two different ways, one that cued an early closure reading and another that

cued late closure. The grammatical difference between the two spoken utterances on each trial was cued by either pitch differences, duration differences, or both pitch and duration differences. Subjects completed three blocks of trials.

Speech in noise threshold

Participants completed a speech in noise test, a full description of the materials and methods for which is described in Boebinger et al., (2015). On each trial, a participant was presented with a short sentence from the BKB corpus (Bench et al., 1979) spoken by a female talker in the presence of competing background voices. Participants reported as much of the speech as they comprehended to the experimenter, who marked how many key words (max of 3) were reported correctly. The test adapted the signal to noise level adaptively with a staircase procedure.

Pitch and Duration Thresholds

Pitch and duration thresholds were obtained with *MLP* (Grassi and Soranzo, 2009), an adaptive thresholding procedure based on the maximum likelihood method. For both the pitch and duration threshold test participants completed 3 blocks of 30 trials. On each trial they heard 3 complex tones in a 3 alternative forced choice design. For the pitch test, participants indicated which of the 3 tones was higher in pitch. For the duration test they judged which of the 3 tones was longer. Responses were made by pressing button 1, 2, or 3 on a keyboard. For each block a threshold was calculated, the point where the probability of a correct response was 33% (chance). For both the

Duration and Pitch tests, the median threshold value across the 3 blocks was taken forward to statistical analysis.

Statistical analysis

The data in the Linguistic Focus and Linguistic Phrase and Musical Phrase tests were analyzed with R. Linear mixed effects models were estimated with the *lme4*, with Group (Amusic or Control), Condition (Pitch, Duration or Both) and their interaction entered as fixed effects, and Item and Subject as random intercepts. P-values for these effects were calculated with likelihood ratio tests of the full model against a null model without the variable in question. Comparisons of predicted marginal means were performed with *lsmeans.*

The dependent variable for the Musical Phrase Test was calculated by identifying the raw response value between -50 and 50 (for each trial) based on the position along the response bar on which the participant clicked, with -50 corresponding to responses on the extreme end of the Incomplete side of the scale. The sign of the data point for Incomplete trials was then inverted so that more positive scores always indicated correct performance and greater scores indicated more accurate categorization of musical phrases.

The dependent variable that went into the model for the Focus and Linguistic Phrase tests was whether each response was CORRECT or INCORRECT. Because the dependent variable was binary, we used the generalized linear mixed models (*glmm)* function in the *lmer* package to estimate mixed effects logistic regressions, and we report odds ratios as a measure of effect size.

Because distribution of pitch thresholds and speech in noise thresholds in the amusic group were non-normal, group comparisons of pitch discrimination, duration discrimination, and speech in noise detection thresholds were assessed non-parametric Mann-Whitney-Wilcoxon tests, and relationships between continuous variables were tested with Kendall's Tau-b.

Results

Musical Phrase Test

Performance scores were affected by type of auditory cue (main effect of Condition $\chi^2(4) = 30.76$, $p < .001$), by whether the participants had amusia (main effect of Group $\chi^2(3) = 9.43$, $p = 0.02$) as well as by the interaction of those factors ($\chi^2(2) = 8.21$, $p = 0.02$). When only pitch cues were present, amusics' performance was significantly lower than controls, and was in fact not different from chance, with the confidence interval for their mean performance score including zero (Figure 4; Table 3). This suggested they were unable to perform the task using cues from pitch alone. Indeed, amusics performed less accurately when relying on pitch cues alone than when they could rely on duration, or on pitch with duration together (Table 1). However, when amusics could rely on duration, either alone, or together with pitch, their performance did not differ from controls. In fact, in the naturalistic Both condition, the mean score for amusics and controls differed by less than 1 response point ($CONTROL_{mean}$ - $AMUSIC_{mean} = 0.70$). Marginal means by group and condition are plotted in Figure 4.

Table 1: Musical Phrase Test, all pairwise contrasts (p-values FDR-corrected)

Condition	Group	Contrast	Est	SE	df	T	p
Both	~	CONT vs AMUS	0.7	1.7	124.78	0.41	0.683
Duration	~	CONT vs AMUS	-1.41	1.7	124.78	-0.83	0.528
Pitch	~	CONT vs AMUS	4.6	1.7	**124.78**	**2.7**	**0.024**
~	CONT	Both vs Duration	3.94	2.82	142.82	1.4	0.298
~	CONT	Both vs Pitch	5.42	2.82	142.82	1.92	0.128
~	CONT	Duration vs Pitch	1.48	1.53	4513.25	0.97	0.5
~	AMUS	Both vs Duration	1.84	2.8	137.67	0.66	0.577
~	AMUS	Both vs Pitch	9.32	2.8	**137.67**	**3.33**	**0.005**
~	AMUS	Duration vs Pitch	7.48	1.48	**4513.25**	**5.06**	**<.001**

Linguistic Focus Test

Overall, both groups performed best when they heard pitch and duration together, worst when only duration cues were present, and in between when there were only pitch cues (main effect of Condition $\chi^2(4) = 168.4$, $p < .001$). This suggests that both groups benefitted from redundant cues, and that pitch was a more useful cue for detecting focus than duration. On the whole, controls performed more accurately than amusics (main effect of Group $\chi^2(3) = 14.63$, $p = 0.002$). However, the two groups were differentially affected by whether pitch or duration cues were present in the stimuli (interaction of Group X Condition $\chi^2(2) = 12.05$, $p = 0.002$). When relying on duration alone, amusics performed similarly to controls, but when they had to rely on pitch alone or pitch together with duration, they performed significantly less accurately. Plots of marginal means by Group and Condition are in Figure 4, and

statistics for all pairwise contrasts (corrected for multiple comparisons) are shown in Table 2.

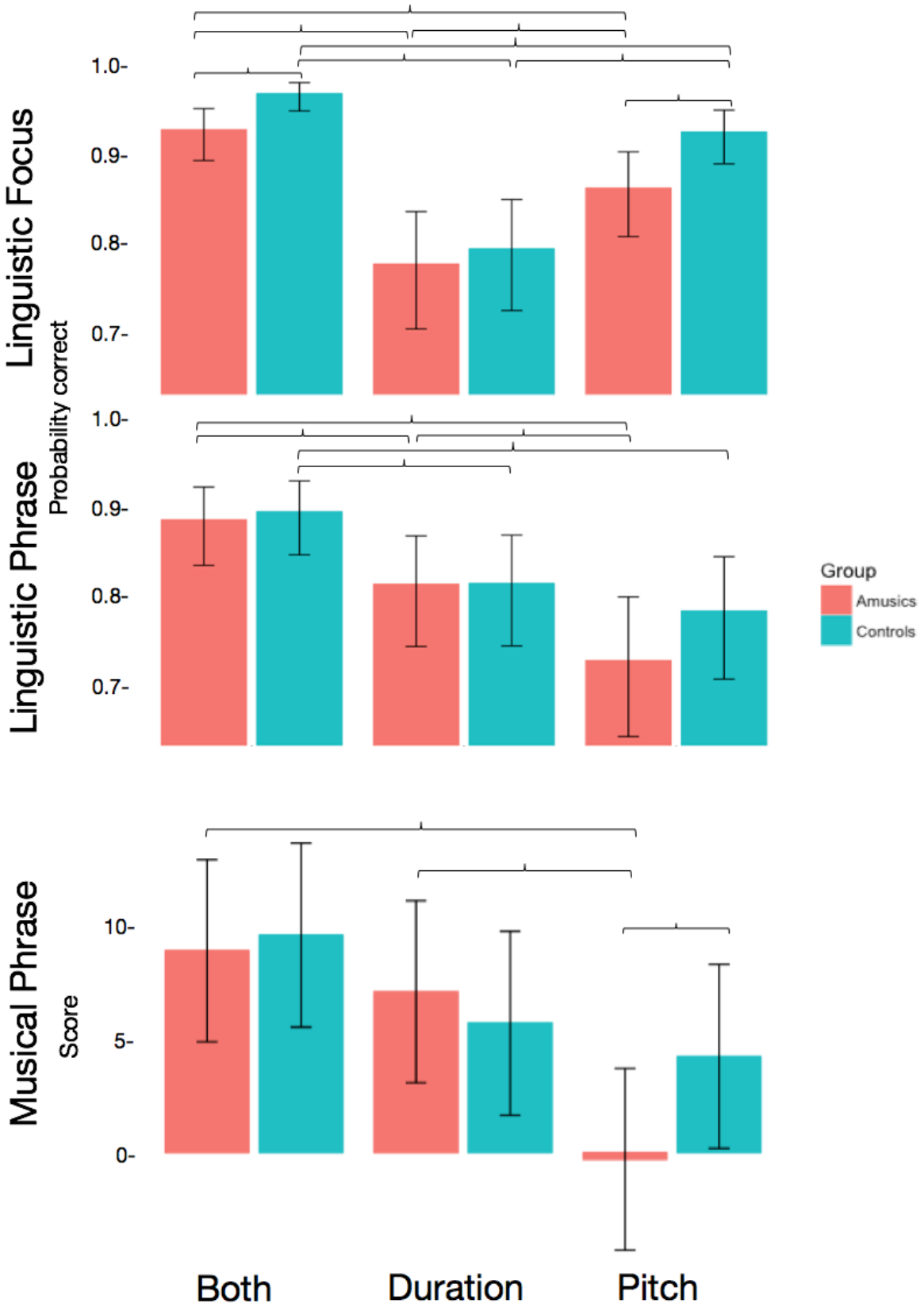

Figure 4. Results of the Linguistic Focus, Linguistic Phrase and Musical Phrase tests. Bars show 95% confidence intervals and brackets indicate significant pairwise contrasts (FDR-corrected).

Table 2: Linguistic Focus test: pairwise comparisons of marginal means (p values FDR corrected).

Condition	Group	Contrast	OR	SE	Z	p
Both	~	CONT vs AMUS	2.44	0.13	**2.71**	**0.009**
Duration	~	CONT vs AMUS	1.11	0.24	0.39	0.697
Pitch	~	CONT vs AMUS	2.00	0.14	**2.39**	**0.019**
~	AMUS	Both vs Pitch	2.06	0.37	**4.01**	**<.001**
~	AMUS	Both vs Duration	3.71	0.64	**7.56**	**<.001**
~	AMUS	Pitch vs Duration	1.80	0.28	**3.83**	**<.001**
~	CONT	Both vs Pitch	2.52	0.62	**3.77**	**<.001**
~	CONT	Both vs Duration	8.15	1.84	**9.31**	**<.001**
~	CONT	Pitch vs Duration	3.23	0.57	**6.65**	**<.001**

Linguistic Phrase Test

Amusics did not differ significantly from controls in overall accuracy (main effect of Group $\chi^2(3) = 2.69$, $p = 0.44$) nor was their performance compared to controls differently affected by which acoustic cues were present (interaction of Group X Condition $\chi^2(2) = 2.33$, $p = 0.31$). Cue type did affect performance (main effect of Condition $\chi^2(4) = 83.06$, $p < .001$). Because there was a Condition difference but no Group differences, we collapsed over Group: participants performed most accurately when both pitch and duration were present, least accurately when they had to rely on pitch cues alone, and in between when they relied on duration alone (Both > Pitch OR = 2.60, SE = 0.28, Z=8.7, p <.001, Both > Duration OR = 1.85, SE = 0.21, Z=5.5,

p<.001; Duration > Pitch OR = 1.41, SE = 0.07, Z = -3.4, p = 0.0017; p-values adjusted using Tukey method). As in the Focus test, redundant cues benefitted both groups, but contrary to the pattern in the Focus Test, duration was a more reliable cue to linguistic phrase boundary perception than pitch.

We hypothesized *a priori* that amusics would rely more on duration than pitch, and this was confirmed (Table 3). For completeness, all other (post-hoc) pairwise comparisons are also reported. All of the contrasts were modest, with odds ratios ranging from 1.01 to 2.88. Results are plotted in Figure 4.

Table 3: Post hoc contrasts, Linguistic Phrase Test (FDR-corrected)

Condition	Group	Contrast	OR	SE	Z	p
Both	~	CONT vs AMUS	1.10	0.26	0.32	0.841
Duration	~	CONT vs AMUS	1.01	0.27	0.02	0.985
Pitch	~	CONT vs AMUS	1.35	0.20	1.12	0.338
~	AMUS	Both vs Pitch	2.88	0.44	**7.00**	**<.001**
~	AMUS	Both vs Duration	1.77	0.28	**3.64**	**0.001**
~	AMUS	Duration vs Pitch	1.63	0.08	**3.56**	**0.001**
~	CONT	Both vs Pitch	2.34	0.37	**5.37**	**<.001**
~	CONT	Both vs Duration	1.93	0.31	**4.10**	**<.001**
~	CONT	Duration vs Pitch	1.21	0.12	1.34	0.268

Pitch, Duration and Speech In Noise Thresholds

Amusics showed higher pitch change detection thresholds than controls (Controls = .21 semitones; Amusics = 0.55 semitones; Mann Whitney Wilcoxon W=29, p<.001) but did not differ from controls in duration thresholds (Controls = 29 ms, Amusics = 32 ms, W = 129, p=.74). In the speech and noise test, the average SNR level visited across the experiment also did not differ between groups (Mean SNR level visited

Controls = -1.52, Amusics = -0.83; Mann-Whitney-Wilcoxon W=155.5, p=.17)(Figure 5).

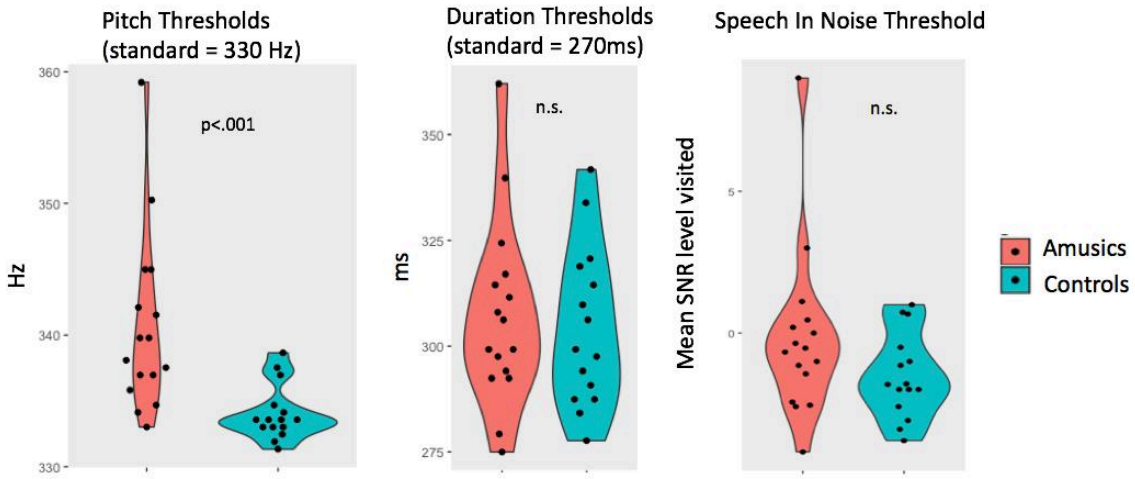

Figure 5: Perceptual thresholds between amusics and controls. Amusics differed between groups in their ability to detect pitch changes. They did not differ in ability to detect duration changes or to hear speech in noise.

Correlations

Next, we examined whether basic auditory processing abilities could explain our results by testing whether individuals' thresholds for detection of pitch change and duration change were correlated with performance scores on any of the conditions of our three main tasks. Pitch psychophysics thresholds were correlated with performance on the Focus Test when both pitch and duration cues were present (r_t = -0.31, p=.026) or when pitch alone was a cue (r_t = -0.3, p=.029). Pitch thresholds were also correlated with Musical Phrase Test performance when pitch cues were the sole cue to phrase endings (r_t = -0.29, p=.023). No significant correlations emerged for duration or speech in noise. Scatter plots are shown in the Supplementary Materials.

Discussion

Music and speech carry redundant acoustic cues, with pitch, duration and amplitude often providing information about the same feature. We tested whether these redundancies make communicative auditory signals robust to individual differences in perceptual abilities. A model population was selected who we confirmed had a deficit for perceiving auditory channel (pitch) but preserved ability for another (duration). We tested how well amusics could perceive musical structures, as well as prosody, when they could only rely on their impaired channel alone, an unimpaired channel alone, or both together. We found that amusics did have difficulty perceiving music and speech based on pitch alone, but that performance improved when they could also rely on a redundant, unimpaired channel. The present work demonstrates that redundancy in the production of communicative signals leads to robustness in perception of such signals, in the face of individual differences in perception.

Musical aptitude is often measured with tests that target specific domains like perception of melody or rhythm (Wallentin et al., 2010; Gorden, 2002). This is, however, unlike actual music listening in the real world. Naturally produced musical structures, such as musical phrases, are often conveyed by simultaneous (i.e. redundant) cues (Krumhansl, 1987). In our musical phrase test, we examined whether our model population could integrate pitch and duration cues in music to make judgments of musical phrases. Indeed, we found that amusics performed poorly relative to controls when they had to rely on pitch alone, but performed as well as controls when they could rely on duration either alone or together with pitch. We thus find no evidence that amusics' intuitions about naturalistic musical phrase structure is impaired and conclude that amusics are able to use duration cues to

parse musical structures. Previous studies of musical phrase judgments have found that duration and pitch cues carry equal weights, without additional benefits from being able to combine the two (Palmer & Krumhansl, 1987). It appears that while amusics may not fully appreciate aspects of music that relate to pitch, they can parse musical structures when another relevant cue is available.

Previous work had suggested that amusics' perception of contrastive pitch accents may be unimpaired (Ayotte et al., 2002; Peretz et al., 2002; Patel et al., 2005). Here we showed a small but statistically significant difference between amusics and controls in the Focus Test for the naturalistic condition, although overall performance was high ($p_{(correct)}$ AMUSIC = 93%, CONTROL = 97%). When participants were forced to rely on pitch alone, the gap between the groups widened ($p_{(correct)}$ AMUSIC = 86%, CONTROL = 93%), and when only duration cues were present the groups performed similarly ($p_{(correct)}$ AMUSIC = 78%, CONTROL = 80%; Figure 4). From this we can conclude that people with a specific auditory deficit are able to compensate to an extent (if not completely) by integrating a cue that is a not a reliable cue for them (pitch) with one that is more reliable (duration).

Similarly to emphatic accents, phrase boundaries are also cued redundantly by pitch and duration. Duration provides important cues to syntactic boundaries through both final lengthening and rhythmic stress placement (Scott, 1982), and pitch provides such cues through (for example) a marked fall in pitch near a one phrase's end, followed by a rise at the start of the next (Cooper and Sorenson, 1977). Previous work indicates that pitch and duration are about equally important cues for phrase perception (Streeter, 1978). Amusic participants performed more accurately on

duration only than *pitch only* trials in the single cue conditions, whereas control participants could rely on either cue about equally. We did not detect any group differences. However, performance on the Both condition was significantly more accurate than on either of the individual cue conditions, across both groups. This suggests that participants could integrate processing across both cue types to achieve higher performance than when they had to rely on either single cue.

One possible outcome, in principle, was that amusics would show superior duration-processing that they had developed to compensate for their pitch deficit. The data here do not support this. The amusics showed similar and not significantly more accurate duration perception ability than controls across the music and language tests, as well as similar psychophysical duration discrimination thresholds. It is possible that, rather than developing exceptional duration processing ability, all that is necessary is a re-weighting in perception to emphasize dimensions where perception is more accurate.

For simplicity's sake we only examined two auditory dimensions -- pitch, where we suspected our groups would show a difference, and duration, where we believed they would not. Outside the laboratory there are other cues that amusics could take advantage of, such as amplitude changes, which are also associated with phrase boundaries and pitch accents (Streeter, 1978; Sluijter & Van Heuven, 1996), although a recent report suggests amplitude processing may also be impaired in amusia (Whiteford & Oxenham, 2017). Accents also carry visual correlates, like head movements, beat gestures, and eyebrow raises (e.g. Beskow et al., 2006; Krahmer

& Swerts, 2007; Flecha-García, 2010), which amusics should also be able to use to compensate for their pitch impairment in audiovisual speech perception. Further research could tease apart the individual contributions of each of these cues.

Furthermore, while our model population was able to integrate pitch and duration together to perform the tasks, it is possible that other groups might have difficulty with this; for instance individuals with autism have difficulty integrating information across multiple senses (Marco et al., 2011). Further work should be done with other groups with known specific auditory difficulties such as adults and children who we would suspect would show impaired temporal but not pitch perception, e.g. those with autism (O'Connor, 2012) or ADHD (Riccio et al., 1994), or beat deafness (Phillips-Silver et al., 2011). When the results are known, the tests could then be adapted for training, by increasing or decreasing the size of pitch and duration cues in the stimuli appropriately. Furthermore, our participants were on average aged over 60 years and therefore have had decades of speech perception experience. It remains to be seen whether children or young adults with perceptual difficulties would also be able to integrate cues in the same way older ones can.

The results of our speech in noise perception test indicated that English amusics are unimpaired relative to controls. However, it is possible that their task strategy may have differed. Further work is needed to determine to what extent amusics are able to use pitch, duration, and amplitude to separate sound streams and maintain attention to a target talker in the context of competing information, and how prosody perception difficulties may be exacerbated by working memory or other task

demands (see Whiteford & Oxenham, 2017). Indeed, naturalistic speech perception generally does not take the form of listening to isolated sentences in a quiet laboratory. The methodological approach we take here (selective morphing of individual auditory variables) could also be applied to longer portions of discourse, such as monologues, to determine how amusic prosodic perception functions under more ecologically valid linguistic, sensory and mnemonic demands.

Our results showcase how communicative systems are adapted for wide audiences in unobvious ways. Perception can, on the surface, appear to be seamless and universal, with most people appearing to arrive at the same interpretations from the same information. This, however, can mask the true diversity of human experience.

Contributions

A.T.T. developed the study concept. All authors contributed to the design. K.J. performed testing, data collection, data analysis and drafted the manuscript. F.D. and A.T.T. provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Stuart Rosen, Marcus Pearce, Laura Staum-Casasanto, Lori Holt, Alex Martin, Aniruddh Patel, Clare Press and Lauren Stewart for helpful comments and discussion. We also thank all our participants. The work was funded by a Wellcome Trust Seed Award #109719/Z/15/Z to A.T.T. and a Reg and Molly Buck Award from SEMPRE to K.J.

References

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: a group study of adults afflicted with a music- specific disorder. *Brain*, *125*(2), 238-251.

Bench, J., Kowal, Å., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British journal of audiology*, *13*(3), 108-112

Beskow, J., Granström, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. In *Ninth International Conference on Spoken Language Processing*.

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., & Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *The Journal of the Acoustical Society of America*, *137*(1), 378-387.

Cooper, W. E., & Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *The Journal of the Acoustical Society of America*, *62*(3), 683-692.

Deguchi, C., Boureux, M., Sarlo, M., Besson, M., Grassi, M., Schön, D., & Colombo, L. (2012). Sentence pitch change detection in the native and unfamiliar language in musicians and non-musicians: Behavioral, electrophysiological and psychoacoustic study. *Brain research*, *1455*, 75-89.

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, *52*(6), 542-554.

Foxton, J. M., Nandy, R. K., & Griffiths, T. D. (2006). Rhythm deficits in 'tone deafness'. *Brain and cognition*, *62*(1), 24-29.

Grassi, M., & Soranzo, A. (2009). MLP: a MATLAB toolbox for rapid and reliable auditory threshold estimation. *Behavior research methods*, *41*(1), 20-28.

Grondin, S. (1993). Duration discrimination of empty and filled intervals marked by auditory and visual signals. *Attention, Perception, & Psychophysics*, *54*(3), 383-394.

Hutchins, S., Gosselin, N., & Peretz, I. (2010). Identification of changes along a continuum of speech intonation is impaired in congenital amusia. *Frontiers in psychology*, *1*.

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., & Yang, Y. (2010). Processing melodic contour and speech intonation in congenital amusics with Mandarin Chinese. Neuropsychologia, 48(9), 2630-2639.

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., & Yang, Y. (2012). Impaired categorical perception of lexical tones in Mandarin-speaking congenital amusics. Memory & cognition, 40(7), 1109-1121.

Kawahara, H., & Irino, T. (2005). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. *Speech separation by humans and machines*, 167-180.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396-414.

Lerdahl, F., & Jackendoff, R. (1985). *A generative theory of tonal music*. MIT press.

Liu, Fang, Aniruddh D. Patel, Adrian Fourcin, and Lauren Stewart. "Intonation processing in congenital amusia: discrimination, identification and imitation." *Brain* 133, no. 6 (2010): 1682-1693.

Liu, F., Jiang, C., Wang, B., Xu, Y., & Patel, A. D. (2015). A music perception disorder (congenital amusia) influences speech comprehension. *Neuropsychologia*, *66*, 111-118.

Lu, Y., & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*(12), 1253-1262

Marco, E. J., Hinkley, L. B., Hill, S. S., & Nagarajan, S. S. (2011). Sensory processing in autism: a review of neurophysiologic findings.

Nan, Y., Sun, Y., & Peretz, I. (2010). Congenital amusia in speakers of a tone language: association with lexical tone agnosia. Brain, 133(9), 2635-2642.

O'Connor, K. (2012). Auditory processing in autism spectrum disorder: a review. *Neuroscience & Biobehavioral Reviews*, *36*(2), 836-854.

Palmer, C., & Krumhansl, C. L. (1987). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(1), 116.

Patel, A. D., Foxton, J. M., & Griffiths, T. D. (2005). Musically tone-deaf individuals have difficulty discriminating intonation contours extracted from speech. *Brain and cognition*, *59*(3), 310-313.

Patel, A. D., Wong, M., Foxton, J., Lochy, A., & Peretz, I. (2008). Speech intonation perception deficits in musical tone deafness (congenital amusia). *Music Perception: An Interdisciplinary Journal*, *25*(4), 357-368.

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing research*, *308*, 98-108.

Peretz, I., Ayotte, J., Zatorre, R. J., Mehler, J., Ahad, P., Penhune, V. B., & Jutras, B. (2002). Congenital amusia: a disorder of fine-grained pitch discrimination. Neuron, 33(2), 185-191.

Peretz, I., Champod, A. S., & Hyde, K. Varieties of musical disorders: the montreal battery of evaluation of amusia. *Ann N Y Acad Sci* 999**,** 58-75 (2003)

Peretz, I., & Vuvan, D. (2016). Prevalence of congenital amusia. *bioRxiv*, 070961.

Phillips-Silver, J., Toiviainen, P., Gosselin, N., Piché, O., Nozaradan, S., Palmer, C., & Peretz, I. (2011). Born to dance but beat deaf: a new form of congenital amusia. *Neuropsychologia*, *49*(5), 961-969.

Riccio, C. A., Hynd, G. W., Cohen, M. J., Hall, J., & Molt, L. (1994). Comorbidity of central auditory processing disorder and attention-deficit hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *33*(6), 849-857.

Schaffrath, H. (1995). The Essen Folksong Collection in Kern Format. [computer database]. D. Huron (ed.). Menlo Park, CA: Center for Computer Assisted Research in the Humanities, 1995.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *The Journal of the Acoustical Society of America*, *71*(4), 996-1007.

Sluijter, A. M., & Van Heuven, V. J. (1996, October). Acoustic correlates of linguistic stress and accent in Dutch and American English. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (Vol. 2, pp. 630-633). IEEE.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *The Journal of the Acoustical Society of America*, *64*(6), 1582-1592.

Vuvan, D. T., Nunes-Silva, M., & Peretz, I. (2015). Meta-analytic evidence for the non-modularity of pitch processing in congenital amusia. *cortex*, *69*, 186-200.

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, *20*(3), 188-196.

Wightman, C. W., Shattuck- Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707-1717.

Whiteford, K. L., & Oxenham, A. J. (2017b). Auditory deficits in amusia extend beyond poor pitch perception. *Neuropsychologia*, *99*, 213-224.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960-967.