Degeneracy makes music and speech robust to individual differences in perception

Kyle Jasmin[1*], Fred Dick[1], Lori Holt[2], and Adam Taylor Tierney[1]

1. Department of Psychological Sciences
Birkbeck, University of London

2. Department of Psychology, Carnegie Mellon University

[*]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 convey structure through spectral and temporal cues. Individuals' abilities to perceive these cues vary widely, and yet most people comprehend music and speech easily. How? Here we investigated whether degeneracy – multiple cues performing the same function – makes music and speech robust to such individual differences. We tested a model population with a severe deficit for perception of pitch but not duration (congenital amusics) and matched controls on speech prosody and music perception tasks. Although amusics were impaired when only pitch cues were available, they perceived speech and music normally when both cues were present. Moreover, in a separate fine-grained cue-weighting prosodic perception task, amusics down-weighted their unreliable channel (pitch) and up-weighted their reliable one (duration) compared to controls. The results suggest that music and speech exploit degeneracy to ensure message transmission, and that individual listeners in turn weight auditory dimensions advantageously across degenerate channels.

Introduction

Auditory communication systems like music and language convey information through relatively continuous sound streams. At an abstract level, however, these streams consist of smaller units (notes, motifs, words) combined hierarchically into larger structures (lines, phrases, sentences [1]). Comprehending structural aspects of these signals requires identifying how adjacent elements (like words in language and 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 dimensions [2-4]. How, then, are communicative auditory signals like music and speech perceived so successfully, despite large individual differences in auditory perception abilities?

The answer may be that musical and speech, like other evolved systems, have specific properties that make them robust to perturbation. One such property is structural *redundancy*, or the presence of multiple identical structures, such as repetitions of the same note sequence in birdsong. Another such property is *degeneracy* -- a term that arose in biological systems theory[5] and has been more recently been applied in animal signalling [6] -- which refers to the presence of multiple *different* structures that perform the *same* overlapping function. It has been theorized that degeneracy of acoustical properties in speech (multiple different acoustic cues performing same linguistic function) could make speech robust to, e.g., background noise [7] and to any isolated perceptual deficits related to speech [8], but to our knowledge this has not been demonstrated empirically. We hypothesized that structural degeneracy could make not just speech, but also music, robust to diverse perceptual abilities. Indeed, in both music and speech, pitch, duration and amplitude changes often co-occur in time, thereby providing multiple cues to the same structural feature. 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 shift toward longer note durations [9]). 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 [10,11]. *Linguistic focus* (emphasis on a word) is also indicated acoustically by a pitch excursion, durational lengthening and an amplitude increase [12].

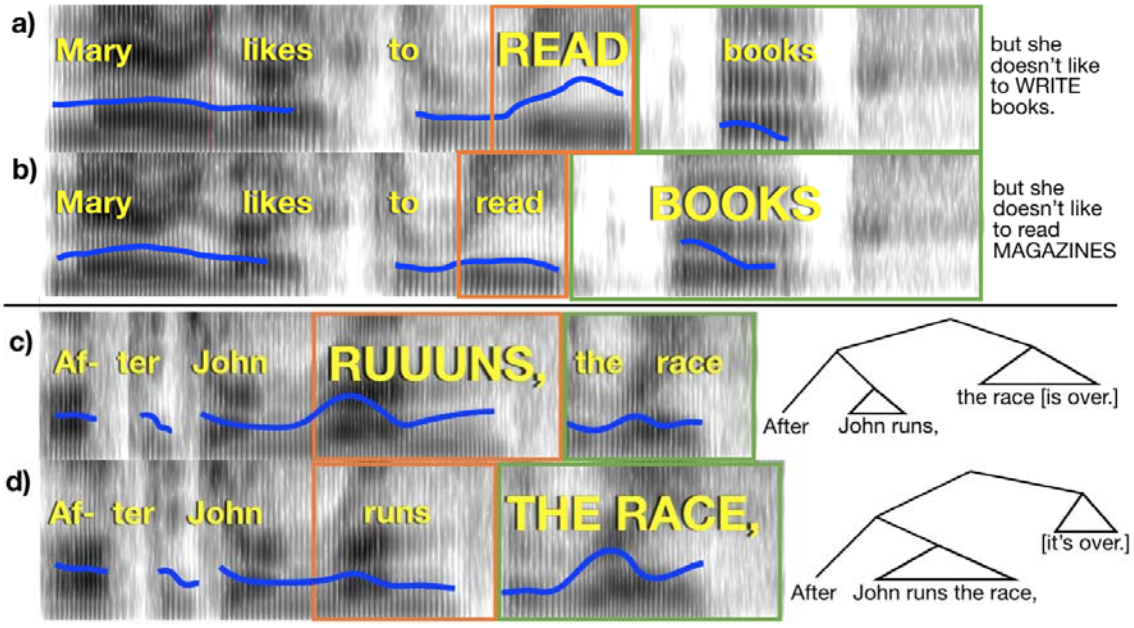

Figure 1. Pitch and duration correlates of emphatic accents and phrase boundaries. Spectrograms of stimuli used in the experiment (time on horizontal axis, frequency on vertical axis, and amplitude in grayscale), with linguistic features cued simultaneously by pitch and duration (the "Combined" condition). Blue line indicates pitch contour. Width of orange 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 is at right. C) a phrase boundary occurs after "runs". D) a phrase boundary occurs after "race". Syntactic trees are indicated at right to illustrate the structure conveyed by acoustics of the stimuli.

Does degeneracy make speech robust to large individual differences in perceptual ability? And does degeneracy also make *music* robust to such individual differences? 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 [13]. Laboratory tests have shown that amusics have difficulty with distinguishing musical melodies based on pitch alone [14]. Amusics sometimes struggle with pitch-related speech tasks [15-see 20 for a meta-analysis], but not invariably [14,21,22]. In real-life situations, amusics may be able to compensate for their impaired pitch perception by relying on degenerate cues to musical and prosodic features. If our model population (with severe deficits) can take advantage of degenerate cues in perceiving speech, this would suggest that individuals with less severe deficits may also do so.

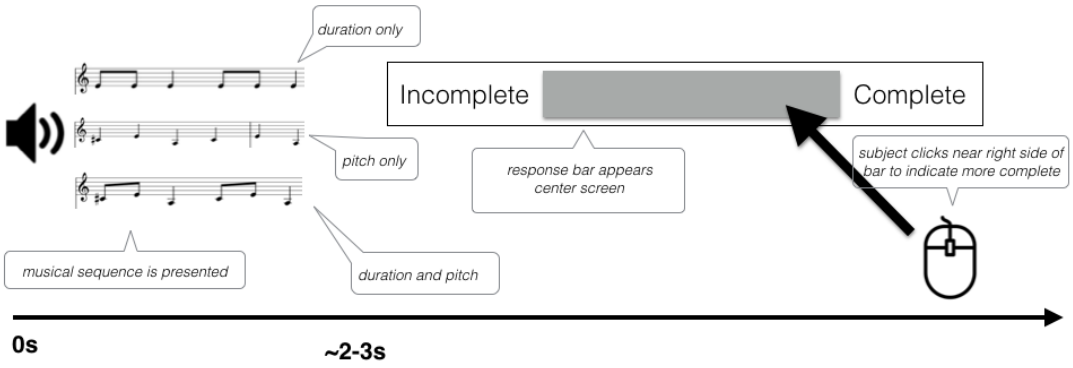

Figure 2. Schematic of trial structure for the Musical Phrase Test. Participants heard a musical sequence that was either a complete musical phrase or straddled a boundary of two musical phrases. They then indicated how complete they thought the phrase sounded by clicking with a mouse at a point along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

In an experiment on music perception (the Musical Phrase Test; Fig 2), 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 of duration, their performance should be improved when they can rely on degenerate cues (pitch and duration), compared to when they must rely solely on an impaired cue (pitch). Next, in two linguistic experiments (Fig. 3), we measured the extent to which subjects used pitch and duration cues to perceive linguistic

focus (or 'emphatic accents; e.g. '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 acoustic manipulation so that participants needed to rely on pitch cues alone, duration cues alone, or could use both combined. Given amusics' near-lack of self-reported language issues (e.g., only 7% reported problems with speech perception in everyday life [23]), we predicted that amusics would perform similarly to controls when they could also take advantage of duration cues (as in natural speech), but more poorly on trials when they had to rely on pitch cues alone.

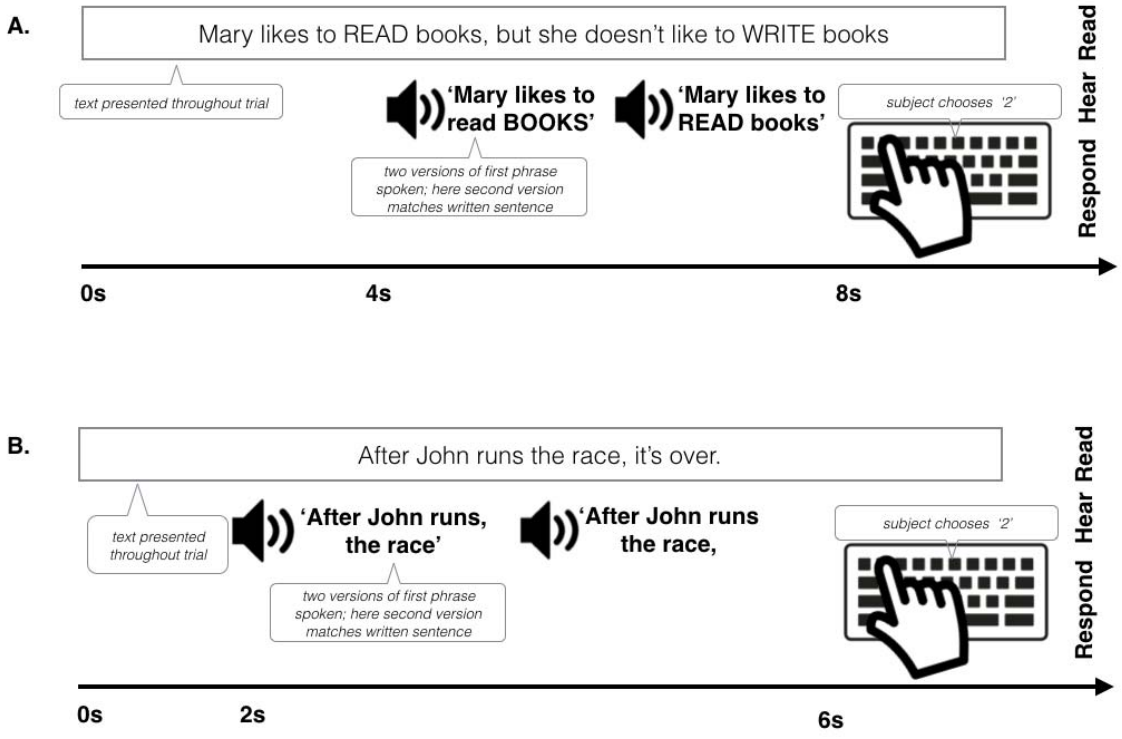

Figure 3: Example trial structure for the linguistic focus test (A) and the linguistic phrase test (B). First, a single sentence was presented visually, and the participants read it to themselves. Next, two auditory versions of the first part of the sentence were played sequentially, only one of which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

As mentioned, some linguistic features are indicated by multiple cues. However, these cues can differ in their *perceptual weight*—that is, one of several cues is often relied upon more than others (i.e. it is *primary* relative to other cues that are *secondary*). In a final experiment, we hypothesized that when individuals are impaired in the perception of a primary cue, they would down-weight it in favor of a secondary cue for which their perception is not impaired. Conversely, if an individual's perception of a secondary cue (but not primary) is impaired, they should have no need to re-weight perceptual cues. To test these predictions, we assessed perceptual cue weighting across a 'prosody space' (within which *pitch* was a primary cue; Fig. 4A) and also a 'phonetic space' (within which *duration* was primary; Fig. 4B). Both stimulus spaces fully crossed manipulations of acoustic pitch and duration such that stimuli indicated one interpretation or another, to varying degrees. Participants with amusia and controls repeatedly categorized tokens sampled from each 2-D acoustic space to provide a measure of the 'perceptual weight' that each acoustic dimension carried in these prosodic and phonetic judgments [24,25]. If amusics take advantage of degeneracy in speech, they should advantageously down-weight a primary cue they have difficulty perceiving in favor of a secondary cue they perceive well.

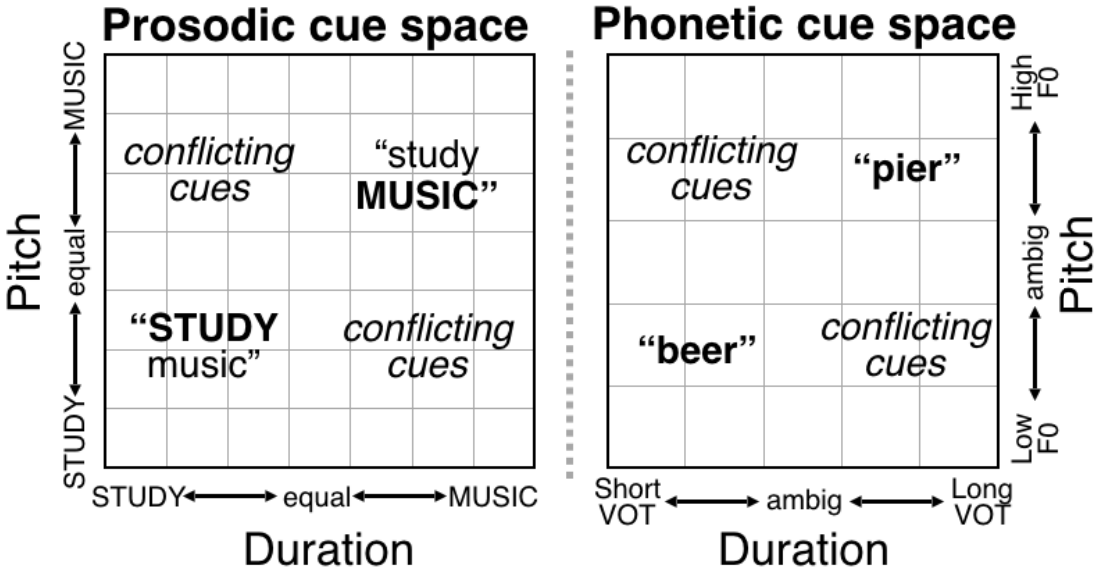

Figure 4: Schematic depiction of Prosodic and Phonetic cues spaces. The prosodic and phonetic cue spaces. Duration cues in the voice that cued an emphasis on STUDY or MUSIC to varying degrees were crossed with cues from pitch. Sometimes pitch and duration were both more likely to cue the same interpretation (upper right and bottom left corners) and other times the cues conflicted (upper left and bottom right). A similar schematic of the phonetic cue space. The initial F0 excursion of the vowel (pitch cue) and voice onset time (duration cue) were crossed to create a phonetic stimulus space [25].

Results

Basic auditory processing

As expected, amusics as a group were less sensitive to pitch differences than controls (Mann Whitney Wilcoxon W (MWWW) = 29, $p < 0.001$), but did not differ from controls in tone duration discrimination (MWWW = 129, $p = 0.74$) or speech-in-noise threshold (MWWW = 155.5, $p = 0.17$; Supplemental Fig. 1).

Musical phrase perception

The musical phrase perception test (schematic in Fig. 2) tested participants' ability to perceive how well a series of notes resembled a complete musical phrase. Auditory cue type affected participants' accuracy in identifying complete versus incomplete phrases, with highest scores when both cues were present, lowest when only pitch was present, and intermediate scores when only duration was present (main effect of Condition $\chi^2(4) = 30.76$, $p < .001$, see Table 1 for pairwise statistics). Compared to controls, amusics wereoverall less accurate (main effect of Group $\chi^2(3) = 9.43$, $p = 0.02$) and also differentially affected by which cue was present (Group x Condition interaction $\chi^2(2) = 8.21$, $p = 0.02$). FDR-corrected pairwise tests showed that when only pitch cues were available, the average amusic's performance was significantly lower than controls (p=.024; Table 1). Indeed, the confidence interval around amusics' mean score included zero (Fig. 5C; Table 1), suggesting that they were unable to perform the task using pitch cues alone. By contrast, when amusics could rely

on duration cues alone, or both pitch and duration together (as in the Combined condition, where cues were present as in naturalistic melodies), amusics and controls did not differ.

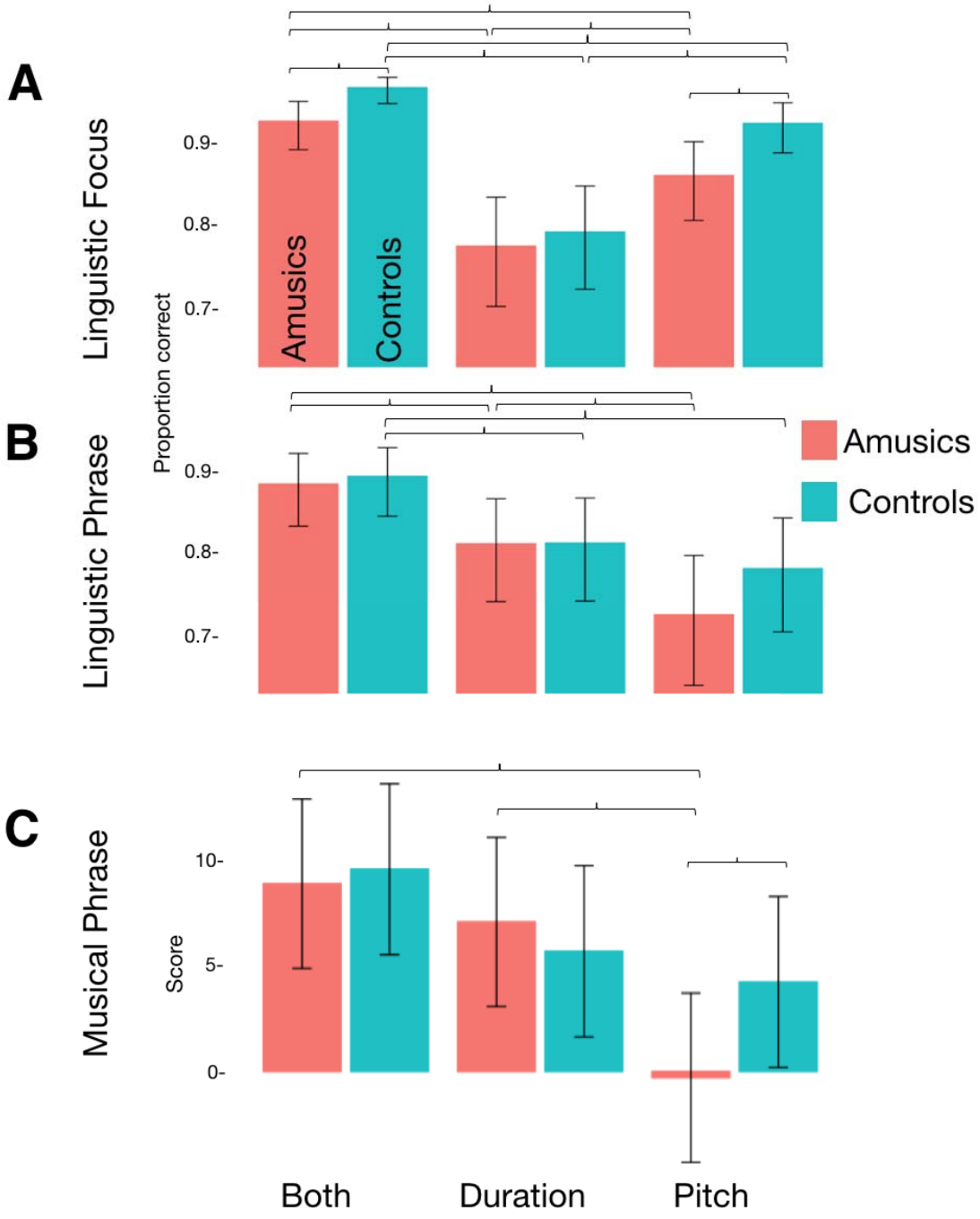

Figure 5. Results of the Linguistic Focus, Linguistic Phrase and Musical Phrase tests. Bars show 95% confidence intervals and brackets indicate significant pairwise contrasts (FDR-corrected).

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

Condition	Group	Contrast	Beta	SE	df	T	p
Combined	~	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	Combined vs Duration	3.94	2.82	142.82	1.4	0.298
~	CONT	Combined 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	Combined vs Duration	1.84	2.8	137.67	0.66	0.577
~	AMUS	Combined 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

The linguistic focus test (schematic Fig. 3A) measured participants' ability to detect where a contrastive accent was placed in a sentence, based on only one type of auditory cue (Pitch or Duration) or both combined (as in natural speech). As shown in Fig. 5A and Table 2, 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 $\chi^2(4) = 168.4$, $p < 0.001$). This suggests that both groups benefitted from degenerate 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 needed to rely on pitch they performed significantly less accurately (p=.019; Table 2).

This disadvantage held where pitch was the sole cue, as well as in the combined pitch+duration cue condition.

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

Condition	Group	Contrast	OR	SE	Z	p
Combined	~	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	Combined vs Pitch	2.06	0.37	**4.01**	**<.001**
~	AMUS	Combined vs Duration	3.71	0.64	**7.56**	**<.001**
~	AMUS	Pitch vs Duration	1.80	0.28	**3.83**	**<.001**
~	CONT	Combined vs Pitch	2.52	0.62	**3.77**	**<.001**
~	CONT	Combined vs Duration	8.15	1.84	**9.31**	**<.001**
~	CONT	Pitch vs Duration	3.23	0.57	**6.65**	**<.001**

Linguistic Phrase Test

The Linguistic Phrase Perception Test (schematic Fig 3b) measured participants' ability to detect phrase boundaries in speech which are cued by pitch only, duration only, or both pitch and duration. Cue type affected performance across groups (main effect of Condition $\chi^2(4) = 83.06$, $p < 0.001$). Participants performed least accurately when they had to rely on pitch cues alone, better when they relied on duration alone, and most accurately when both pitch and duration were present together (see Fig. 5B and Table 3). As in the Focus test, degenerate 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.

Amusics did not differ significantly from controls in overall accuracy (main effect of Group $\chi^2(3) = 2.69$, $p = 0.44$) nor were the groups' performance significantly differently affected by which acoustic cues were present (interaction of Group X Condition $\chi^2(2) = 2.33$, $p = 0.31$). Because we had hypothesized *a priori* that amusics would rely more on duration

than pitch (and that controls would show similar performance across the two conditions), we conducted pairwise contrasts to test this prediction. Amusics did indeed show significantly greater accuracy with duration than with pitch cues (p=.001; Table 3), whereas controls did not. (For completeness, all other (post-hoc) pairwise comparisons are also reported).

Table 3: Post hoc contrasts, Linguistic Phrase Test

Condition	Group	Contrast	OR	SE	Z	P
Combined	~	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	Combined vs Pitch	2.88	0.44	**7.00**	**<0.001**
~	AMUS	Combined vs Duration	1.77	0.28	**3.64**	**0.001**
~	AMUS	Duration vs Pitch	1.63	0.08	**3.56**	**0.001**
~	CONT	Combined vs Pitch	2.34	0.37	**5.37**	**<0.001**
~	CONT	Combined vs Duration	1.93	0.31	**4.10**	**<0.001**
~	CONT	Duration vs Pitch	1.21	0.12	1.34	0.268

Perceptual Weighting in Prosody and Phonetic categorization

These tasks tested a subset of the amusic (N=11) and control (N=11) participants on prosodic (e.g., a phrase perceived either as 'Dave likes to STUDY music' versus 'Dave likes to study MUSIC') and phonetic (words perceived as 'beer' or 'pier') categorization across a 2-dimensional acoustic spacse that fully crossed pitch and duration dimensions, as shown in Fig. 4. From prior literature, it was known that durational cues (VOT) are primary to pitch cues (F0) for perceiving phonetic voicing [25,26], whereas the results of focus experiment in the present work, as well as prior literature [27,28] indicated that pitch was a more informative cue than duration for perception of prosodic accents. We hypothesized that amusics - who are demonstrably less sensitive to pitch than controls - would accordingly ‘down-weight’ pitch as

a cue (relative to controls) while 'up-weighting' duration, for the prosodic space where they were relatively less sensitive to the primary dimension. (For calculation of normalized cue weights see Methods).

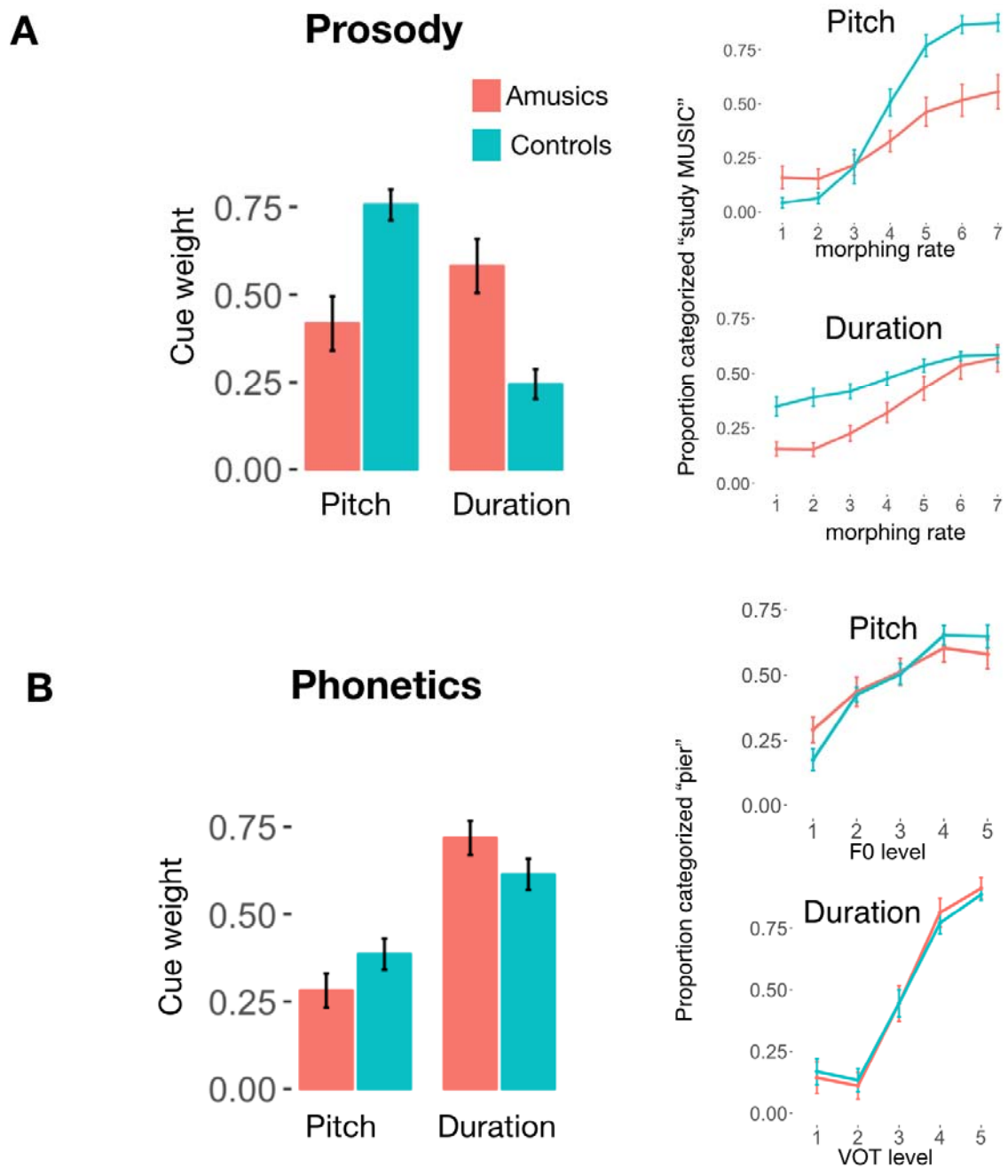

Figure 6: Comparison of pitch and duration cue weights for prosody and phonetic perception. A) Left: Mean cue weights plotted by group and condition (left). Mean categorization response plotted at each level of pitch, collapsed over duration; and each level

of duration collapsed over pitch. **B)** Analogous plots for the phonetic categorization. Bars indicate standard error of the mean.

Prosodic space results indicated that perceptual weights for the pitch dimension were higher for controls than amusics, while duration weights were higher for amusics than controls (group comparisons $T_{20} = 3.81$, p=.001), suggesting that amusics and controls did indeed rely upon these two dimensions differently during the task (Fig. 6A). T-tests were used for the main group comparisons because normalizing the pitch and duration cue weights relative to each other (so they sum to 1) causes them to be non-independent, and therefore a Group X CueType interaction test was inappropriate. However, we note that an ANOVA on the raw cue weights (before normalization) further confirmed this pattern (interaction of CueType X Group, $F(1,39 = 10.3$, $p = 0.002$; full analysis of raw cue weights in Supplement). To gain a finer-grained understanding of each group's weighting, in Figure 7A we plotted mean responses for both groups at each of the 49 locations in the stimulus space depicted schematically in Figure 4. We then compared the matrices cell-by-cell (two-sample T-tests Controls>Amusic, FDR-corrected, Fig. 7B). As noted, pitch and duration sometimes cued the same response and sometimes conflicted, as seen in the schematic of the stimulus set (Fig 4.) All (corrected) significant differences that emerged were in the top half of the matrix, where pitch cued an emphasis on “MUSIC”. Most (12 out of 16) of these significant group differences occurred in the 16 stimuli of the upper-left quadrant, where emphasis was placed on “STUDY” by duration cues, and “MUSIC” by pitch cues. In this “conflicting” quadrant of the stimulus space, where pitch and duration pointed to differing interpretations, amusics relied on the duration cues to make their response more often than controls did. Finally, we calculated a metric reflecting participants’ relative ability to discriminate pitch and duration in complex tones (see Methods), and tested whether this metric explained subjects’ relative preference to rely on pitch or duration in the prosody task. Participants with finer pitch than

duration discrimination thresholds tended to have higher pitch (than duration) cue weights (Kendall Tau-b r=0.43, p=.005; Fig. 8A).

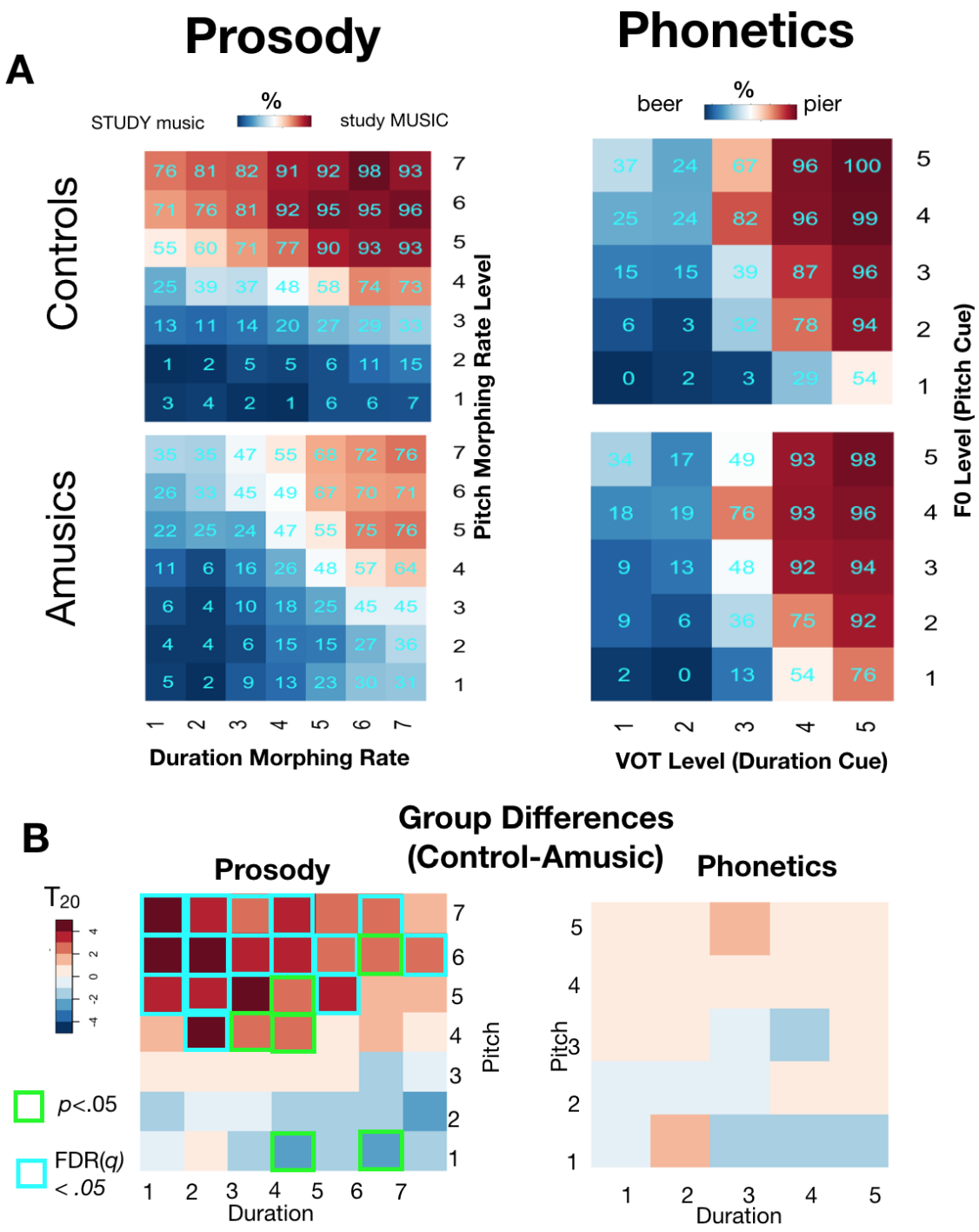

Figure 7: When pitch and duration cues conflict, amusics rely on duration. A) Heatmaps indicate proportion of trials categorized as "study MUSIC" (for the prosody portion, panel A) or "pier" (phonetics portion, panel B), for the Control and Amusic groups. **B)** Group

difference (Controls – Amusics) heatmaps T-statistics.. When duration and pitch conflicted in the prosody task (duration indicated emphasis on STUDY, but pitch indicated emphasis on MUSIC; upper-left quadrant of stimulus space), amusic participants chose the duration-based response more often than controls. Teal outlines indicate significant group differences (corrected for multiple comparisons). Uncorrected results (p<.05) are indicated with green outlines.

Unlike the results from prosodic perception, where pitch was the primary cue and amusics showed a different relative weighting of pitch and duration, no such significant difference was detected for phonetic judgments where duration was primary (group effect T_{20} = 1.58, p = 0.13). This difference between the group effects, across experiments, was confirmed statistically (interaction of Group X Experiment, $F_{(1,40)}$ = 4.44, p=.04). Although the mean responses between amusics and controls did not differ, the results of the t-test at each stimulus location are presented in Fig. 7B for comparison. No results survived even an uncorrected threshold of p < 0.05. Relative ability to discriminate tone pitch vs duration also did not correlate with performance (Fig. 8B). Full results of each individual subject, from both experiments, are plotted in Fig. 9.

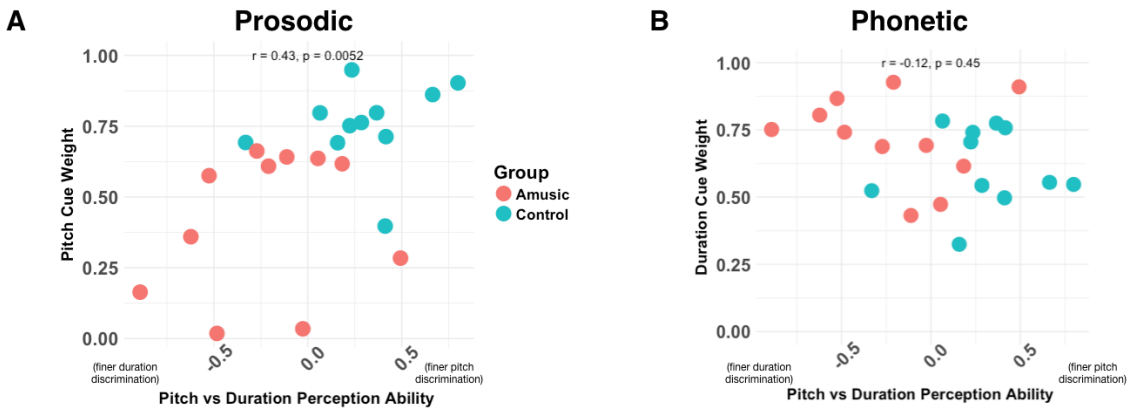

Figure 8: Correlations between pitch vs duration perceptual discrimination and cue weights. A. Individual with finer pitch relative to duration discrimination thresholds tended

to have higher normalized pitch cue weights in the prosodic categorization task. Correlation shown is Kendall's Tau-b. **B.** Similar plot for the phonetic experiment, with pitch vs duration threshold plotted against duration cue weights (with no significant effect).

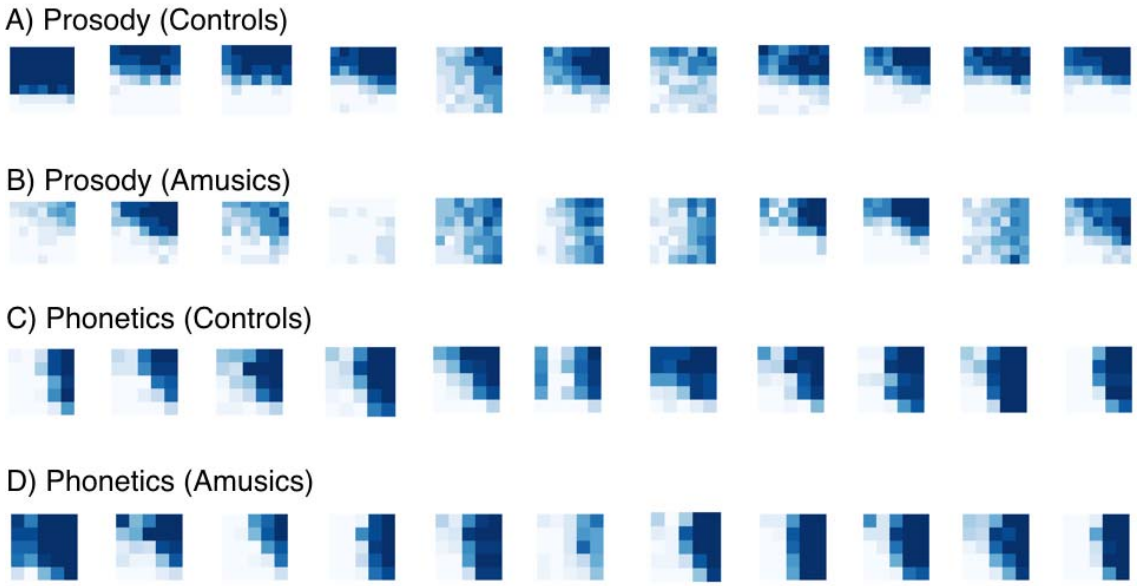

Figure 9: Individual heatmaps indicating responses for all subjects for the prosody blocks (A-B) and phonetic blocks (C-D). Horizontal axis indicates duration and vertical axis indicates pitch. Plots in A,C and in B,D present the participants in the same order (e.g. data from leftmost plot in A is from the same participant as leftmost plot in C).

Discussion

Music and speech carry multiple acoustic cues, with pitch, duration and amplitude often providing information about the same feature. We tested whether this property might make communicative acoustic signals more robust to individual differences in perceptual abilities. A model population was selected who we confirmed had a deficit in perceiving one acoustic dimension (pitch) but preserved ability for another (duration). We tested how well amusics could perceive musical and linguistic structure disambiguated by their impaired channel (pitch) alone, an unimpaired channel (duration) alone, or both together. We found that amusics had difficulty extracting structure in music and speech based on pitch alone, but that performance improved when they could also rely an unimparied channel, either on its own, or together with pitch. We further demonstrated that amusics listened differently from controls:

when perceiving linguistic focus (to which the primary cue was pitch), amusics up-weighted their more reliable channel (duration) over their unreliable one (pitch). No such re-weighting occurred for the phonetic task (for which duration was already the primary cue). The present work demonstrates that degeneracy in the production of communicative signals leads to robust categorization in the face of individual differences in auditory perception.

Prior studies of perceptual categorization of speech have generally examined how the "average listener" weights multiple acoustic cues during speech perception. As discussed, such studies have found that certain dimensions are "primary" for a specific task because they provide reliable information about speech categories and are, therefore, weighted highly in perception. Other, less reliable, "secondary" dimensions are also present, but are only used when primary dimensions are ambiguous. For example, in linguistic focus, pitch is generally primary and duration is secondary [27,28]. However, individual differences in dimensional weighting have been reported [29-32] and the sources of these differences have been unclear. Here we identified one potential source of these differences: the perceptual fidelity of acoustic dimensions for individual listeners. Utilizing an extreme case, amusia, we demonstrated that listeners' perceptual weight across acoustic dimensions is impacted by individual differences in auditory perceptual fidelity of pitch. This suggests that, in speech categorization, an individual may weight a cue more heavily because she processes the corresponding auditory dimension more robustly. Similarly, listeners who process a given feature at a coarser grain may direct auditory attention away from it, decreasing its perceptual weight relative to other features.

However, several results in the current study suggest that such a simplistic interpretation - where 'pitch' and 'duration' are conceptualized as single auditory dimensions - does not provide a full account of the data. For instance, while amusics do, on average, show less precise discrimination of pitch differences between two complex tones than controls,

there is not only a good degree of overlap in pitch thresholds between the groups, but the detection threshold in even the least pitch-sensitive amusic participant was well below that required to detect the informative pitch changes in any of the speech and music tasks. That said, the significant correlation between tone pitch vs duration discrimination ability and cue weights (see Fig. 8) and the small correlations between pitch psychophysics thresholds and performance on the music, linguistic phrase, and linguistic focus tasks with disambiguating pitch information (see Supplement) suggest that sensitivity to pitch does inform its use in higher-level tasks.

In the phonetic categorization task, amusics showed entirely typical use of pitch cues, weighting their decisions based on rapid pitch changes that differed by less than two semitones across levels. However, the same participants showed much less sensitivity to a greater and more prolonged pitch deviation in the context of the prosodic and musical tasks (on the order of the better part of an octave). We have suggested that the reason showed different cue weights for the prosodic but not phonetic experiment is because pitch is a primary cue for linguistic focus but only a secondary cue for the phonetic contrast (voicing). An alternative explanation, however, is that these findings may reflect a specific amusic deficit for integration of pitch information across longer time frames, rather than a simple encoding deficit. Indeed, in the prosody experiment, pitch information unfolded over time in the order of seconds, whereas in the phonetic experiment the pitch excursion was over milliseconds.

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. The present data suggest that rather than

developing exceptional duration processing ability, all that may be necessary is a re-weighting in perception to emphasize dimensions where perception is more accurate.

Musical aptitude is often measured with tests that target specific domains like perception of melody or rhythm [33,34]. 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. degenerate) cues [9]. Here, we 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 [9], a finding we replicated in our control participants. Amusics, on the other hand, showed a gain from degeneracy. While amusics may not fully appreciate aspects of music that relate to pitch, we show that they can parse musical structures when another relevant cue is available. Future work should investigate whether this spared musical perception in amusics extends to other musical features which are communicated by degenerate cues, such as musical beat perception [35].

For perception of phrase boundaries, pitch and duration are about equally important [10]. Performance on the Combined condition in the phrase task was significantly more accurate than on either of the individual cue conditions, suggesting that participants (even non-amusic ones) integrated across both cue types to achieve higher performance than when they had to rely on either single cue.

To keep the experimental design simple, 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 individuals could take advantage of, such as vowel quality, which is also associated with phrase boundaries and pitch accents [10,12]. Accents also carry visual correlates, such as head movements, beat

gestures, and eyebrow raises [e.g. 36-38], which individuals may also be able to use to compensate for their pitch impairment in audiovisual speech perception. Further research could examine the individual contributions of each of these cues.

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 [39], ADHD [40], or beat deafness [4]. Our model population was able to integrate pitch and duration together to perform the tasks, but it is possible that other groups might have difficulty with this; for instance individuals with autism have difficulty integrating information across multiple senses [41].

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.

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. Participants were compensated £10 per hour of participation. 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). The amusia status for 15 amusics and 10 controls had been determined previously and they had taken part in other psychological studies of amusia. The rest were recruited especially for this study via an online MBEA test. Amusia is a rare condition with 1.5% prevalence [13]. The sample size was therefore limited by our ability to recruit, screen and test qualifying participants.

Musical Phrase Perception Test

Stimuli

The stimuli consisted of 100 musical phrases taken from a corpus of folk songs [42]. They appeared in three conditions: Combined – 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 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 (Combined Cues, then Duration Only, then Pitch Only).

Linguistic focus perception task

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 recorded these sentences as they were spoken by an actor who placed contrastive 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 [43], 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.

Linguistic phrase perception test

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 [44] 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 Standard Southern British English-speaking male (trained as an actor) recorded early and late closure versions of the sentences in his own standard Southern English dialect. 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 combined. 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 3B). 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.

Cue Weighting Experiment

An additional experiment assessed the extent to which participants used fundamental frequency (F0) and duration cues to make phonetic and prosodic judgments. The consonant /p/ differs phonetically from /b/ both in voice onset time and fundamental frequency [45]. Likewise (as already discussed) linguistic focus is often cued both via a pitch accent and a durational lengthening. We recruited participants from the original set of amusic and control participants with 11 amusic (6F, age = 59.3) and 11 control (8F, age = 60.4) participants volunteering to return for the experiment.

Stimuli

Each phonetic block consisted of repetitions of a single word, spoken by a female American English speaker, that varied from “beer” (IPA: /bier/) to “pier” (IPA: /pier/) along two continua. The voice onset time varied from -5 ms to 15 ms in 5 ms increments. The initial F0 varied from 200 to 320 in 30 Hz increments. Shorter VOT and lower initial F0 sounded more like /b/ while longer VOT and higher initial F0 sounded like /p/ [25]. Each of the 5 F0 (pitch) levels was crossed with each of the 5 VOT (duration) levels to make 25 combinations.

For the focus blocks, the stimulus consisted of the phrase “Dave likes to STUDY music” or “David likes to study MUSIC”, which was excerpted from a voice-morphed item from the Focus Test described above. The duration and pitch information disambiguating the 'focused' word varied from 0% to 100% morphing rates for pitch and for duration, in 17% increments (0%, 17%, 33%, 50%, 67%, 83%, 100%), such that each of the 7 pitch levels

occurred with each of the 7 duration levels to make 49 combinations. Because the dimensions were fully crossed, some combinations of pitch and duration cued an interpretation jointly, others conflicted, and tokens near the center of the space were designed to be more ambiguous (Fig. 4). Examples of the stimuli are provided in online materials.

Procedure

Participants completed this experiment online from home, and were instructed to use headphones. Two subjects did not have computers with headphones and were therefore tested in the lab. The experiment began with instructions to listen to each item and classify whether the initial consonant was “B” or “P” (for the phonetic task) or whether the emphasis resembled “STUDY music” or “study MUSIC”. Responses were made by clicking with their mouse one of two buttons positioned near the center of the screen (“B”/”STUDY music” on left; “P”/”study MUSIC” on right). Each phonetic block contained 50 repetitions of the item (2 measurements at each of the 25 combination of F0 and duration); and each focus block contained 49 repetitions (1 measurement at each of the 49 F0 and duration combinations). The participant completed 20 blocks of stimuli – 10 phonetic blocks (500 trials total) and 10 focus blocks (490 trials total) which alternated, with short breaks interspersed. The entire experiment lasted approximately 60 minutes.

Statistical analysis

Data were analyzed with R. For the musical phrase, linguistic focus and linguistic phrase tests, linear mixed effects models were estimated using *lme4*, with Group (Amusic or Control), Condition (Pitch, Duration or Combined) 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 was entered 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.

For the cue weighting studies, Cue Weights were calculated by constructing a multiple logistic regression for each participant (separately for phonetic and prosodic components) with Pitch and Duration as factors (on integer scales from 1-7 for the prosody component, and 1-5 for the phonetic component, according to the number of stimulus increments). The coefficients estimated from these models were then normalized such that the Pitch and Duration weights summed to one. A large coefficient for, e.g., pitch relative to duration indicated that the pitch factor in the stimulus space explained more variance in participants' categorization judgments than the duration factor. To test for group effects, these Cue Weights for pitch and duration were extracted for each subject and subjected to a T-test .

Because distribution of pitch thresholds were non-normal relationships between pitch and duration thresholds and cue weights were tested with Kendall's Tau-b. The metric indicating relative pitch vs duration discrimination ability was calculated by first subtracting each subject's pitch and duration threshold from the standard used in the psychophysics test (330Hz for pitch; 270ms for duration), then dividing by the standard deviation to obtain a

standard score. The standard scores for pitch and duration were then combined with an asymmetry ratio [(Duration – Pitch) / (Duration + Pitch)] such that higher values indicated finer pitch than duration thresholds, whereas lower values indicated the reverse.

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., A.T.T. and L.H. provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Stuart Rosen, Marcus Pearce, Laura Staum-Casasanto, 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., a Reg and Molly Buck Award from SEMPRE to K.J., and a Leverhulme Trust Early Career Fellowship to K.J.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Lerdahl, F. & Jackendoff, R. *A Generative Theory of Tonal Music*. (MIT Press, 1985).
2. Grondin, S. Duration discrimination of empty and filled intervals marked by auditory and visual signals. *Perception & Psychophysics* **54,** 383–394 (1993).
3. Deguchi, C. *et al.* 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 (2012).
4. Phillips-Silver, J. *et al.* Born to dance but beat deaf: A new form of congenital amusia. *Neuropsychologia* **49,** 961–969 (2011).
5. Price, C. J. & Friston, K. J. Degeneracy and cognitive anatomy. *Trends in Cognitive Sciences* **6,** 416–421 (2002).
6. Hebets, E. A. *et al.* A systems approach to animal communication. *Proc. R. Soc. B* **283,** 20152889 (2016).
7. Winter, B. Spoken language achieves robustness and evolvability by exploiting

degeneracy and neutrality. *BioEssays* **36,** 960–967 (2014).

8. Patel, A. D. Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research* **308,** 98–108 (2014).
9. Palmer, C. & Krumhansl, C. L. Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance* **13,** 116–126 (1987).
10. Streeter, L. A. Acoustic determinants of phrase boundary perception. *The Journal of the Acoustical Society of America* **64,** 1582–1592 (1978).
11. Wightman, C. W., Hufnagel, S. S., Ostendorf, M. & Price, P. J. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America* **91,** 1707–1717 (1992).
12. Sluijter, A. M. C. & van Heuven, V. J. Acoustic correlates of linguistic stress and accent in Dutch and American English. in **2,** 630–633 (IEEE, 1996).
13. Peretz, I. & Vuvan, D. T. Prevalence of congenital amusia. *European Journal of Human Genetics* **25,** 625–630 (2017).
14. Peretz, I. *et al.* Congenital amusia: a disorder of fine-grained pitch discrimination. *Neuron* **33,** 185–191 (2002).
15. Patel, A. D., Wong, M., Foxton, J., Lochy, A. & Peretz, I. Speech intonation perceptuion deficits in musical tone deafness (congenital amusia). *Music Perception: An Interdisciplinary Journal* **25,** 357–368 (2008).
16. Hutchins, S., Gosselin, N. & Peretz, I. Identification of Changes along a Continuum of Speech Intonation is Impaired in Congenital Amusia. *Front. Psychology* **1,** (2010).
17. Nan, Y., Sun, Y. & Peretz, I. Congenital amusia in speakers of a tone language: association with lexical tone agnosia. *Brain* **133,** 2635–2642 (2010).
18. Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J. & Yang, Y. Processing melodic contour and speech intonation in congenital amusics with Mandarin Chinese. *Neuropsychologia* **48,** 2630–2639 (2010).
19. Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J. & Yang, Y. Impaired categorical perception of lexical tones in Mandarin-speaking congenital amusics. *Memory & Cognition* **40,** 1109–1121 (2012).
20. Vuvan, D. T., Nunes-Silva, M. & Peretz, I. Meta-analytic evidence for the non-modularity of pitch processing in congenital amusia. *CORTEX* **69,** 186–200 (2015).
21. Ayotte, J., Peretz, I. & Hyde, K. Congenital amusiaA group study of adults afflicted with a music-specific disorder. *Brain* **125,** 238–251 (2002).
22. Patel, A. D., Foxton, J. M. & Griffiths, T. D. Musically tone-deaf individuals have difficulty discriminating intonation contours extracted from speech. *Brain and Cognition* **59,** 310–313 (2005).
23. Liu, F., Patel, A. D., Fourcin, A. & Stewart, L. Intonation processing in congenital amusia: discrimination, identification and imitation. *Brain* **133,** 1682–1693 (2010).
24. Holt, L. L. & Lotto, A. J. Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America* **119,** 3059–3071 (2006).
25. Idemaru, K. & Holt, L. L. Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance* **37,** 1939–1956 (2011).
26. Francis, A. L., Kaganovich, N. & Driscoll-Huber, C. Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America* **124,** 1234–1251 (2008).
27. Breen, M., Fedorenko, E., Wagner, M. & Gibson, E. Acoustic correlates of information structure. *Language and Cognitive Processes* **25,** 1044–1098 (2010).

28. Chrabaszcz, A., Winn, M., Lin, C. Y. & Idsardi, W. J. Acoustic Cues to Perception of Word Stress by English, Mandarin, and Russian Speakers. *J Speech Lang Hear Res* **57,** 1468–1479 (2014).
29. Hazan, V. & Rosen, S. Individual variability in the perception of cues to place contrasts in initial stops. *Perception & Psychophysics* **49,** 187–200 (1991).
30. Idemaru, K., Holt, L. L. & Seltman, H. Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America* **132,** 3950–3964 (2012).
31. Schertz, J., Cho, T., Lotto, A. & Warner, N. Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics* **52,** 183–204 (2015).
32. Kim, D., Clayards, M. & Goad, H. A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics* **67,** 1–20 (2018).
33. Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C. & Vuust, P. The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences* **20,** 188–196 (2010).
34. Gordon, E. E. *Primary Measures of Music Audiation*. (2002).
35. Hannon, E. E., Snyder, J. S., Eerola, T. & Krumhansl, C. L. The Role of Melodic and Temporal Cues in Perceiving Musical Meter. *Journal of Experimental Psychology: Human Perception and Performance* **30,** 956–974 (2004).
36. Beskow, J., Granström, B., Conference, D. H. N. I.2006. Visual correlates to prominence in several expressive modes. *Ninth International Conference on Spoken Language Processing* (2006).
37. Krahmer, E. & Swerts, M. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language* **57,** 396–414 (2007).
38. Flecha-García, M. L. Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication* **52,** 542–554 (2010).
39. O'Connor, K. Auditory processing in autism spectrum disorder: A review. *Neuroscience and Biobehavioral Reviews* **36,** 836–854 (2012).
40. Riccio, C. A., Hynd, G. W., Cohen, M. J., Hall, J. & Molt, L. Comorbidity of Central Auditory Processing Disorder and Attention-Deficit Hyperactivity Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry* **33,** 849–857 (1994).
41. Marco, E. J., Hinkley, L. B. N., Hill, S. S. & Nagarajan, S. S. Sensory Processing in Autism: A Review of Neurophysiologic Findings. *Pediatric Research* **69,** 48R–54R (2011).
42. Schaffrath, H. & Park, D. H. M. *The Essen folksong collection in kern format.[computer database]*. (1995).
43. Kawahara, H. & Irino, T. in *Speech Separation by Humans and Machines* 167–180 (Kluwer Academic Publishers, 2005). doi:10.1007/0-387-22794-6_11
44. Kjelgaard, M. M. & Speer, S. R. Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity. *Journal of Memory and Language* **40,** 153–194 (1999).
45. Haggard, M., Ambler, S. & Callow, M. Pitch as a Voicing Cue. *The Journal of the Acoustical Society of America* **47,** 613–617 (1970).