Temporal Structure Perception in Tonal and Atonal Music: An Online Behavioural Study

Seung-Goo Kim[1,2,a,*], Daniel Müllensiefen[3,b], Ying Yu[1,4], Tobias Overath[1,c]

[1]Duke University, North Carolina, U.S.A.

[2]Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

[3]University of Hamburg, Hamburg, Germany

[4]Columbia University Vagelos College of Physicians and Surgeons, New York, U.S.A.

[a]https://orcid.org/0000-0003-0558-8547

[b]https://orcid.org/0000-0001-7297-1760

[c]https://orcid.org/0000-0003-4150-2704

[*]Correspondence: seung-goo.kim@ae.mpg.de

April 21, 2026

Keywords: musical structure, 12-tone serialism

Abstract: How does tonal structure shape the perception of temporal information in music? We addressed this question across five online behavioural experiments (N = 487). Using a dedicated scrambling algorithm ('quilting'), segments of tonal (Bach) and atonal (Krenek) violin recordings were rearranged while preserving variable temporal structures (60–3840 ms), and participants rated perceived naturalness. Naturalness increased with the intact segment length for both corpora following a piecewise linear function, with comparable slopes and elbow points—but a consistently higher plateau for tonal music. Tonal and temporal structures thus seem to contribute independently and additively to naturalness perception. A cross-domain comparison with speech quilts revealed a reversed familiarity effect, suggesting domain-specificity for processing in temporal processing. In addition, musicians showed heightened sensitivity specifically in atonal music. In non-musicians, self-reported Percep-

tual Abilities (Gold-MSI) were robustly associated with the tonal music elbow point and the tonal–atonal plateau difference, replicated via cross-validation. These findings demonstrate parallel processing of tonal and temporal structure in music, with musicality shaping temporal integration even without formal training.

1 Introduction

1.1 Temporal and Tonal Structures of Music

Music is an art form that unfolds over time, and much of its emotional power rests on the listener's capacity to anticipate what comes next. A widely accepted account of musical emotion holds that expectation and its violation (e.g., tension and resolution through harmonic progressions) are the primary drivers of affective responses to music (Huron, 2006; Meyer, 1957). Computational models have formalised this intuition: probabilistic n-gram models have been developed to capture melodic expectation (IDyOM; Pearce, 2018) and harmonic expectation (PPM; Harrison et al., 2020), quantifying the information content of each event (either a note or a chord) in a sequence and the uncertainty of a sequence as to what comes next. Empirical evidence confirms that the information content and uncertainty of melody and chord sequences jointly shape listeners' emotional responses (Cheung et al., 2019; Gold et al., 2019). Building on this, an adaptation of predictive coding theory (Friston & Kiebel, 2009) proposes that precision-weighted prediction errors on temporal sequences play a critical role in evoking musical pleasure (Vuust et al., 2022).

Central to all these accounts is the concept of tonal structure: the hierarchical organisation of pitches in Western music, in which certain tones function as stable anchors for perception while others create tension by gravitating towards them. Tonal structure provides the scaffolding upon which melodic and harmonic expectations are built; it is widely present not only in Western European music but across many musical cultures (Mehr et al., 2019). Without a tonal centre, the predictive relationships that underpin musical expectation are fundamentally disrupted.

1.2 Temporal Integration as a Perceptual Mechanism

For expectation-based accounts of musical emotion to operate, listeners must integrate information across time—accumulating enough context to form a prediction before the next event arrives. Temporal integration refers to this process of combining auditory information that arrives closely in time. It is the

mechanism by which a continuous acoustic stream is parsed into discrete musical units such as notes, chords, phrases, and sections.

Evidence from psychoacoustics suggests that the auditory system does not integrate over a fixed window, but instead operates flexibly across a range of timescales (Dau et al., 1997). Neuroscientific studies suggest that, while the fundamental mechanism of temporal integration is the integration of membrane potential within a single neuron, the complex structure of human neural networks gives rise to multi-scale temporal integration across the hierarchy of the auditory system (Norman-Haignere, Long, et al., 2022a; Overath et al., 2012).

This flexibility has a perceptual correlate: listeners are sensitive to the amount of temporal context available when judging the naturalness or coherence of an auditory sequence. When temporal structure is disrupted beyond a certain timescale (e.g., by randomly rearranging short segments of a sound), listeners perceive the result as less natural, and the perception of decreasing naturalness scales with the length of the disrupted segments (Overath et al., 2015). Critically for the present study, this relationship between segment length and perceived naturalness provides a behavioural window into the timescales over which listeners integrate musical information.

1.3 Prior Work

The scrambling paradigm—in which segments of an auditory stimulus are rearranged to disrupt temporal structure beyond a chosen timescale—has been employed in both behavioural and neuroimaging research to probe how listeners process temporal structure in music. Early behavioural work demonstrated that listeners are sensitive to the disruption of long-range musical structure, and that this sensitivity is modulated by musical training and familiarity with Western tonal conventions (Tillmann & Bigand, 1996, 2001). Subsequent neuroimaging studies extended this approach to examine the neural correlates of musical temporal processing: Levitin and Menon, 2003 used short scrambled segments (∼250–350 ms) to show that intact music recruits frontotemporal networks beyond the auditory cortex, while Farbood et al., 2015 employed musicologically-defined boundaries at the levels of the measure (∼1.29 s), phrase (∼6.32 s), and section (∼38.28 s) to demonstrate the involvement of medial prefrontal cortex in processing large-scale musical structure.

However, these studies share two important limitations. First, neither systematically examined the short-timescale range (60–1000 ms) that spans the transition from sub-beat to multi-beat integration: precisely the range where individual notes, chords, and rhythmic groups are formed. Second, neither study controlled for tonal structure: by using exclusively tonal music, they could not distinguish whether observed

effects reflected sensitivity to temporal structure per se or to the disruption of tonal relationships that co-vary with temporal context. It therefore remains unclear how tonal structure and temporal integration interact in shaping the perceptual experience of music.

1.4 Present Study

The present study addresses these limitations through a series of five behavioural experiments with 487 online participants. We ask whether, and at what timescales, tonal structure modulates the perception of temporal structure in music. To this end, we compare listeners' naturalness ratings of tonal music (J. S. Bach's Violin Sonatas and Partitas) and a carefully matched atonal control stimuli (E. Krenek's Sonata for Solo Violin No. 2)—a 12-tone serial work that preserves the stylistic and instrumental characteristics of the tonal stimuli while eliminating any tonal centre. This contrast allows us to disentangle the perceptual contribution of tonal structure from that of temporal structure.

Temporal structure is manipulated using a refined music quilting algorithm (Overath et al., 2015) that rearranges segments of a specified length while minimising acoustic artefacts at segment boundaries. Seven segment lengths, logarithmically spaced from 60 to 3840 ms, were used to systematically vary the amount of preserved temporal context across sub-beat to multi-beat timescales. This approach generates highly realistic scrambled stimuli and allows us to estimate, for each listener, the timescale at which temporal integration critically shapes the perception of musical naturalness.

Four research questions guide the study:

- (RQ1) Does tonal structure influence the perception of musical temporal structure?
- (RQ2) Is temporal integration in music governed by a domain-general mechanism, such that it correlates with temporal integration in speech?
- (RQ3) Does formal musical training modulate temporal integration?
- (RQ4) More broadly, does musical sophistication in non-musicians interact with temporal perception?

Together, these questions address both the specificity of musical temporal processing and its relationship to individual differences in musical experience. All stimuli, de-identified data, and analysis code are publicly available at Code Ocean (https://codeocean.com/capsule/5141283/).

2 Experiment I: Discovery

The first experiment was designed to discover the relationship between the amount of preserved temporal information (i.e., segment lengths) and overall subjective perception (i.e., "naturalness") in representative tonal and atonal music.

2.1 Materials and Methods

2.1.1 Musical Stimuli

Music quilts of tonal and atonal music were created from recordings of J. S. Bach's Violin Sonatas and Partitas and E. Krenek's Sonata for Solo Violin No. 2, both performed by Christoph Schickedanz. Krenek's piece was selected for its absence of a tonal centre while sharing stylistic similarities with the Bach recordings (see Supplementary Methods S1.1).

The quilting algorithm (Overath et al., 2015) rearranges segments of a specified length to disrupt temporal structures exceeding that length while minimising acoustic artefacts at boundaries. Here, the algorithm was further refined to achieve global minimisation of artefacts across all possible initialisations (see Supplementary Methods S1.2). Quilts of 11.52 s duration were generated at seven logarithmically spaced segment lengths: 60, 120, 240, 480, 960, 1820, and 3640 ms—spanning approximately 0.1 to 6.6 beats given average beat durations of 552 ms (Bach) and 608 ms (Krenek).

From the full set of generated quilts, tonal–atonal pairs closely matched in overall spectral and spectrotemporal modulation energy were selected (Norman-Haignere & McDermott, 2018, see Supplementary Methods S1.3). For each of the 14 conditions (2 tonality $\times$ 7 segment lengths), the 4 best-matched exemplar sets constituted Music Set A. All stimuli were loudness-normalised to $\sim$31 sones (ISO 532-1; Supplementary Methods S1.4), cosine-ramped over the initial and final 20 ms, and exported as 44.1 kHz / 16-bit `FLAC` files. Two reference stimuli—an original Bach excerpt ("Original") and matched synthetic noise ("Synth"; Norman-Haignere & McDermott, 2018)—served as naturalness anchors.

2.1.2 Participant Recruitment

Participants were recruited from Amazon Mechanical Turk (MTurk) in July 2021. The inclusion criteria included (a) normal hearing, (b) a high approval rate (exceeding 98%), (c) substantial experience with over 100 approved MTurk tasks, (d) residence in North America (the US or Canada), and (e) use of

headphones for the task. Headphone use was verified through a behavioural screening test using anti-phase stimuli (Woods et al., 2017).

2.1.3 Procedure

After providing informed consent online, participants completed a ∼1-min headphone screening (compensated US$0.01). Those who passed were directed to the main experiment, hosted on the Qualtrics platform at Duke University.

In the main experiment, participants rated the perceived naturalness of each stimulus on a continuous slider (0–10, 1,001 steps; "0 = least, 10 = most natural"). Naturalness was operationalised as the perceptual opposite of the quilting manipulation. Two practice trials (Bach excerpts with the shortest and longest segment lengths: 60 and 3840 ms) preceded the main blocks to familiarise participants with the scale.

The main experiment consisted of four blocks of 16 trials each, comprising 7 Bach and 7 Krenek quilts across all segment lengths, plus two reference stimuli ("Original" and "Synth"). Each condition appeared once per block, repeated across blocks with different exemplars. Trial and block orders were randomised. Participants were required to listen to each stimulus in full (11.52 s) before rating; an optional 30-s break was offered between blocks.

Following the main blocks, participants completed the three Gold-MSI subscales (Active Engagement, Perceptual Abilities, Musical Training; Müllensiefen et al., 2014), along with questions on Western music familiarity and demographic information (i.e., age group, gender group, and ethnicity group). Sessions lasted approximately 25 min; all completed responses were manually reviewed and compensated at US$4.17.

2.1.4 Data Analysis

Of 109 downloaded responses, post-screening flagged suspicious responses based on four criteria: (a) rating "Synth" higher than any other stimulus, (b) adjusting the slider before the 3840-ms segment had fully played, (c) submitting ratings more than 30 s post-stimulus-onset, and (d) straight-lining on >77% of Gold-MSI items. Flagged responses were manually reviewed; 6 were excluded, leaving 103 for analysis.

Although the reference stimuli ("Original" and "Synth") served as reliable group-level naturalness anchors, individual-level reference ratings frequently fell outside the range of other ratings, making normalisation liable to amplify noise (see Supplementary Methods S1.4). Reference ratings were therefore used

only to estimate ceiling and floor confidence intervals, not for individual normalisation.

An elbow function (Overath et al., 2015) was fit to each participant's naturalness ratings as a function of segment length:

$$f(x; a, b, c) = \begin{cases} a(x-b)+c & \text{if } x < b \\ c & \text{if } x \geq b \end{cases} \tag{1}$$

where $x \in \{1, 2, \ldots, 7\}$ indexes segment length as $60 \times 2^{(x-1)}$ ms, $a \geq 0$ is the slope, $b \in [1, 7]$ the elbow point, and $c \geq 0$ the plateau height. A positive slope indicates that the naturalness rating increases as longer temporal structure is intact. The elbow point indicates where this effect saturates: preserving even longer temporal structure does not further contribute to the perception of naturalness. Model fit was assessed by adjusted R^2; the elbow function substantially outperformed a linear fit (Supplementary Methods S1.6).

The main effect of segment length was assessed by testing the slope against zero using one-sample one-sided t-test. The main effect of tonality was assessed by comparing the plateau height between Bach and Krenek quilts using paired-sample two-sided t-test. The interaction between tonality and segment length was tested by comparing slope and elbow point between Bach and Krenek. The family-wise error rate of multiple testing was corrected using false discovery rate (FDR; $M = 11$; Benjamini and Hochberg 1995). Visualisations were produced using custom MATLAB scripts with a `ggplot2`-style colourmap from GRAMM (Morel, 2018).

2.2 Results

2.2.1 Sample Demographics

Demographic details of analysed participants in all experiments are presented in Supplementary Results S2.1. Briefly, among 103 participants, a substantial portion was between 30 and 44 years old (45%), of European descent (71%), and male (55%). Mean Gold-MSI subscales were 32.8 for Active Engagement, 44.6 for Perceptual Abilities, and 15.7 for Musical Training. Notably, these group averages were markedly lower than the normative means (41.5, 50.2, 26.5; unequal-variance two-sample t-test, $P < 10^{-7}$) as reported in the original study based on the UK population (Müllensiefen et al., 2014).

Table 1: Experiment I: fitted parameters of the elbow function ($N = 103$). Units of parameters: Slope, an increase in Naturalness rating over a unit increase in Segment Length steps. Elbow, Segment Length step; Height, Naturalness rating, Adjusted R^2, arbitrary unit. Abbreviations: Std. Dev, standard deviation; CI, confidence interval.

Contrast	Parameter	Mean	Std. Dev.	95% CI	t[102]	P_{FDR}
Bach vs. null	Slope [rating/step]	1.032	0.573	$[0.920, 1.144]$	18.285	$< 10^{-32}$
	Elbow [step]	5.687	1.232	$[5.446, 5.928]$	46.835	$< 10^{-69}$
	Height [rating]	7.646	1.221	$[7.407, 7.885]$	63.548	$< 10^{-82}$
	Adjusted R^2	0.573	0.477	$[0.480, 0.666]$	12.188	$< 10^{-20}$
Krenek vs. null	Slope [rating/step]	1.017	1.237	$[0.775, 1.259]$	8.344	$< 10^{-12}$
	Elbow [step]	5.342	1.717	$[5.006, 5.677]$	31.580	$< 10^{-53}$
	Height [rating]	5.554	1.515	$[5.258, 5.850]$	37.216	$< 10^{-59}$
	Adjusted R^2	0.417	0.505	$[0.318, 0.515]$	8.369	$< 10^{-12}$
Bach vs. Krenek	ΔSlope [rating/step]	0.015	1.388	$[-0.256, 0.286]$	0.109	0.913
	ΔElbow [step]	0.345	2.154	$[-0.076, 0.766]$	1.628	0.117
	ΔHeight [rating]	2.092	1.391	$[1.820, 2.364]$	15.263	$< 10^{-27}$

2.2.2 Naturalness Ratings Increased over Segment Lengths, Interacting with Tonality

Table 1 and Figure 1 provide a comprehensive summary of descriptive and inferential statistics. The elbow function proved to be an effective model for the data (mean adjusted $R^2 = 0.495$). Estimated slopes were consistently positive ($P_{\mathrm{FDR}} < 10^{-12}$) and elbow points were greater than the shortest segment length ($P < 10^{-53}$), indicating that naturalness ratings increased logarithmically with segment length. When comparing estimated coefficients for Bach and Krenek quilts, slopes and elbow points did not differ ($P_{\mathrm{FDR}} > 0.117$). However, the plateau height was significantly higher for Bach than Krenek by 2.092 naturalness rating units ($P_{\mathrm{FDR}} < 10^{-27}$), indicating that naturalness ratings increased similarly across segment lengths for both corpora but reached different plateau levels, with Bach stimuli achieving higher naturalness ratings overall.

2.3 Discussion

Experiment I investigated how temporal structures in music are perceived and how they are modulated by tonal structure. Naturalness ratings were well described by a piecewise linear elbow function. Estimated parameters indicated that the perception of temporal structure was largely similar for tonal and atonal music in terms of both the rate of increase and the saturation (i.e. elbow) point, with no discernible interaction between tonality and segment length. However, the different plateau levels indicate that tonal music was perceived as more natural than atonal music across all segment lengths, reflecting a main effect of tonality.

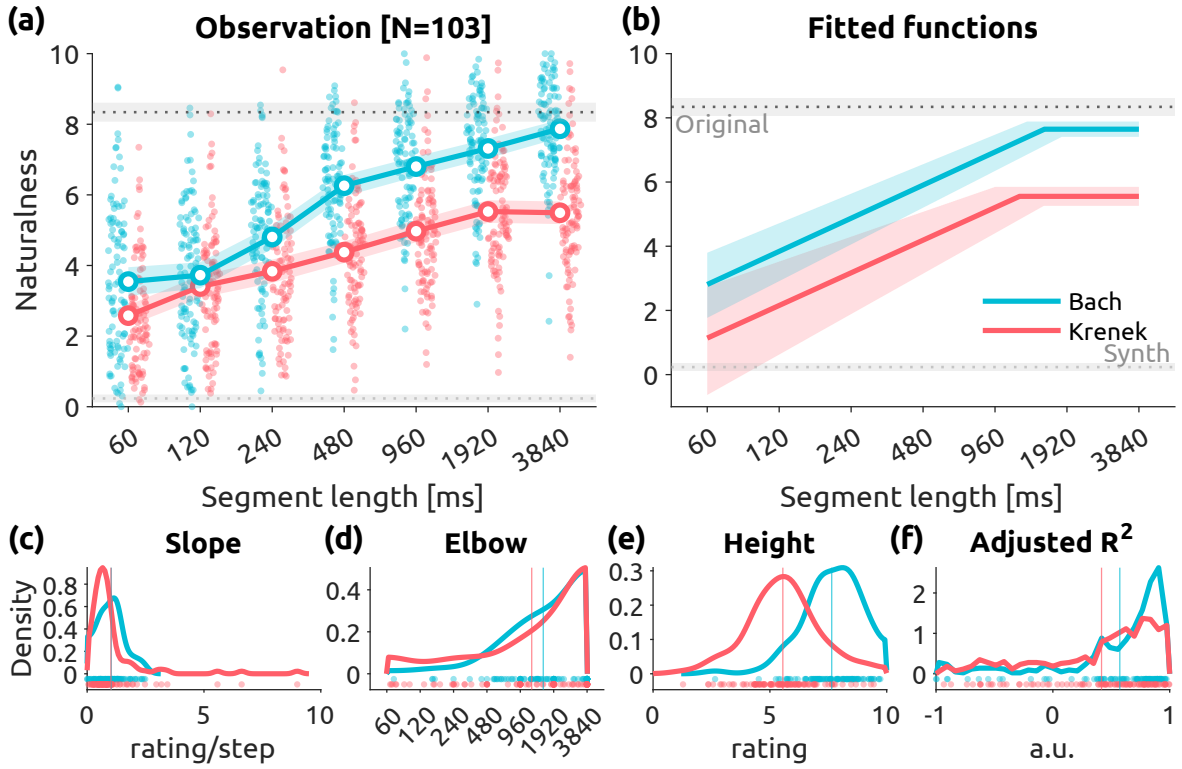

Figure 1: Experiment I: observed means and distribution of fitted parameters. (a) Observed means with 95% confidence intervals [CIs] shaded for Bach [blue] and Krenek [red]. Gray horizontal dotted lines indicate observed means for Original (upper) and Synth (lower) stimuli, also with 95% CIs shaded. (b) Fitted functions with 95% CIs shaded. Raincloud plots [kernel densities and scatter plots] for (c) slope, (d) elbow, (e) height, and (f) adjusted R2. For interpretability, elbow parameters are displayed in milliseconds. Vertical lines in colours denote respective means.

This finding is particularly noteworthy given that the online participants, who reported musical sophistication below normative means, nonetheless exhibited a clear perceptual distinction between tonal and atonal music. To validate these findings, Experiment II was conducted with an independent sample of participants.

3 Experiment II: Replication

In order to validate the reliability of the findings, we replicated Experiment I with independent participants. The materials, stimuli, and procedure were identical to those used in Experiment I.

3.1 Results

One hundred and eight participants were recruited from MTurk in August 2021. Four participants were excluded for overlap with Experiment I and ten were excluded during post-screening. The remaining 94 participants were included in the subsequent analysis.

3.1.1 Sample Demographics

The demographic distribution of participants did not differ from that of Experiment I in terms of age group, gender group, and ethnicity group (χ^2-test, $P > 0.17$). Among 94 participants, a majority were again 30 and 44 years old (62%), of European descent (77%), and male (55%).

3.1.2 Naturalness Ratings Increased over Segment Lengths, Interacting with Tonality

As shown in Table 2 and Figure 2, model fit was comparable to Experiment I (mean adj. $R^2 \geq 0.398$). Consistently with Experiment I, estimated parameters indicated a logarithmic increase in naturalness ratings over segment length ($P_{\mathrm{FDR}} < 10^{-13}$). When comparing Bach and Krenek, only the plateau heights differed ($P_{\mathrm{FDR}} < 10^{-19}$), as in Experiment I.

Table 2: Experiment II: fitted parameters of the elbow function ($N = 94$).

Contrast	Parameter	Mean	Std. Dev.	95% CI	t[93]	P_{FDR}
Bach vs. null	Slope [rating/step]	1.008	0.641	$[0.877, 1.139]$	15.259	$< 10^{-25}$
	Elbow [step]	5.377	1.335	$[5.104, 5.651]$	39.058	$< 10^{-58}$
	Height [rating]	7.616	1.251	$[7.360, 7.873]$	59.034	$< 10^{-73}$
	Adjusted R^2	0.445	0.584	$[0.325, 0.565]$	7.389	$< 10^{-10}$
Krenek vs. null	Slope [rating/step]	0.872	0.944	$[0.678, 1.065]$	8.951	$< 10^{-13}$
	Elbow [step]	5.347	1.652	$[5.009, 5.685]$	31.385	$< 10^{-50}$
	Height [rating]	5.753	1.782	$[5.388, 6.118]$	31.304	$< 10^{-50}$
	Adjusted R^2	0.398	0.512	$[0.294, 0.503]$	7.552	$< 10^{-10}$
Bach vs. Krenek	ΔSlope [rating/step]	0.136	1.105	$[-0.090, 0.363]$	1.196	0.258
	ΔElbow [step]	0.030	1.837	$[-0.346, 0.407]$	0.160	0.874
	ΔHeight [rating]	1.863	1.541	$[1.547, 2.179]$	11.721	$< 10^{-19}$

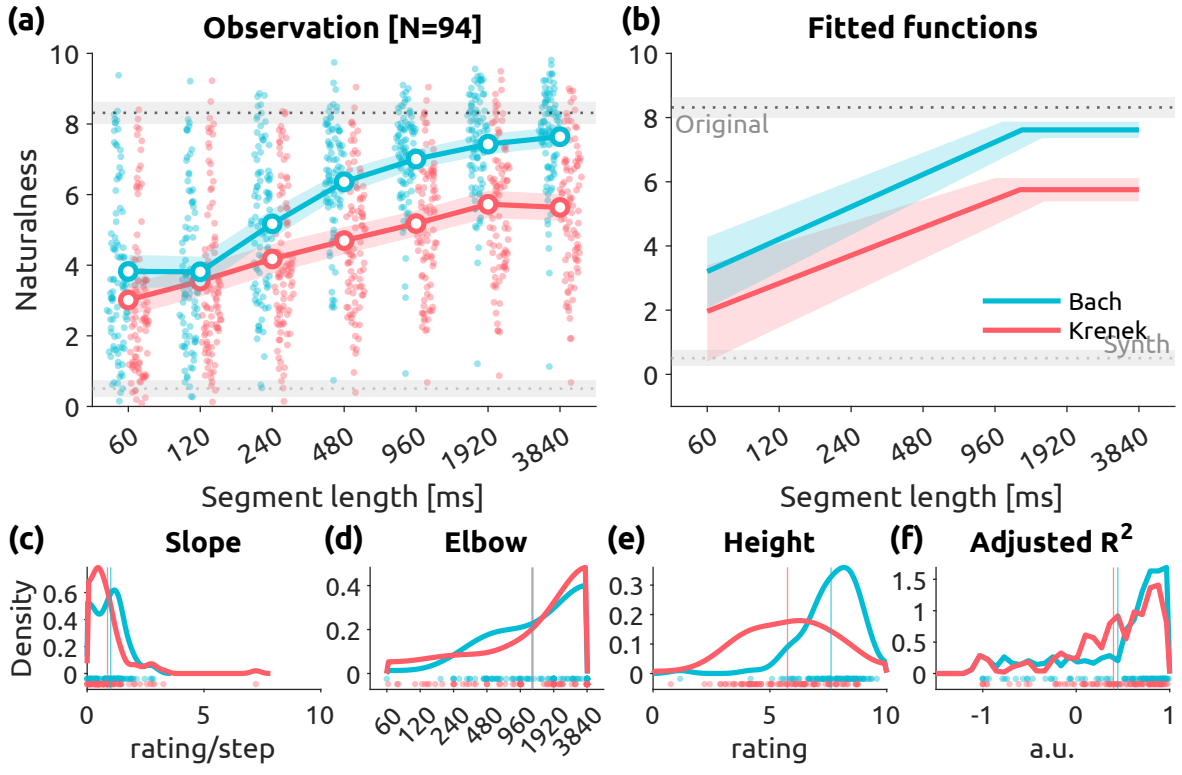

Figure 2: Experiment II: observed means and distribution of fitted parameters. Data are displayed in the identical scheme as in Figure 1.

3.2 Discussion

Consistent with Experiment I, Experiment II replicated strong main effects of segment length and tonality. In this independent sample, which shared similar demographics and musical sophistication levels with Experiment I, the effects of segment length (slopes and elbows) were similar for Bach and Krenek, with a clear and parallel height difference across segment lengths. We therefore investigated whether these effects generalise to different stimulus exemplars.

4 Experiment III: Generalisation

In this experiment, we sought to replicate the main findings with different exemplars. The materials and procedure were identical to those used in Experiments I and II. The stimuli comprised the remaining four exemplars from the 8 top-matched exemplars (Supplementary Methods S1.3).

4.1 Results

Independent participants with no overlap with previous experiments were recruited from MTurk in July 2021. Of 123 participants, 10 were excluded during post-screening. The remaining 108 participants were included in the subsequent analysis.

4.1.1 Sample Demographics

The demographic distribution of Experiment III did not differ from that of Experiment I in terms of age group, gender group, and ethnicity group (χ^2-test, min $P = 0.137$). The majority of analysed 108 participants were between 30 and 44 years old (49%), of European descent (72%), and male (61%).

4.1.2 Naturalness Ratings Increased over Segment Lengths, Interacting with Tonality

Table 3 and Figure 3 display the descriptive and inferential statistics. As in Experiments I and II, model fit was similar for Bach quilts (mean adj. $R^2 = 0.530$). However, for Krenek quilts the fit was worse (mean adj. $R^2 = 0.196$), presumably due to the high rating at 60 ms.Nonetheless, a clear difference in plateau heights (1.468 naturalness rating units higher for Bach than Krenek) was consistently found ($P_{\mathrm{FDR}} < 10^{-22}$). Unlike Experiments I and II, moderate differences in slopes and elbow points were

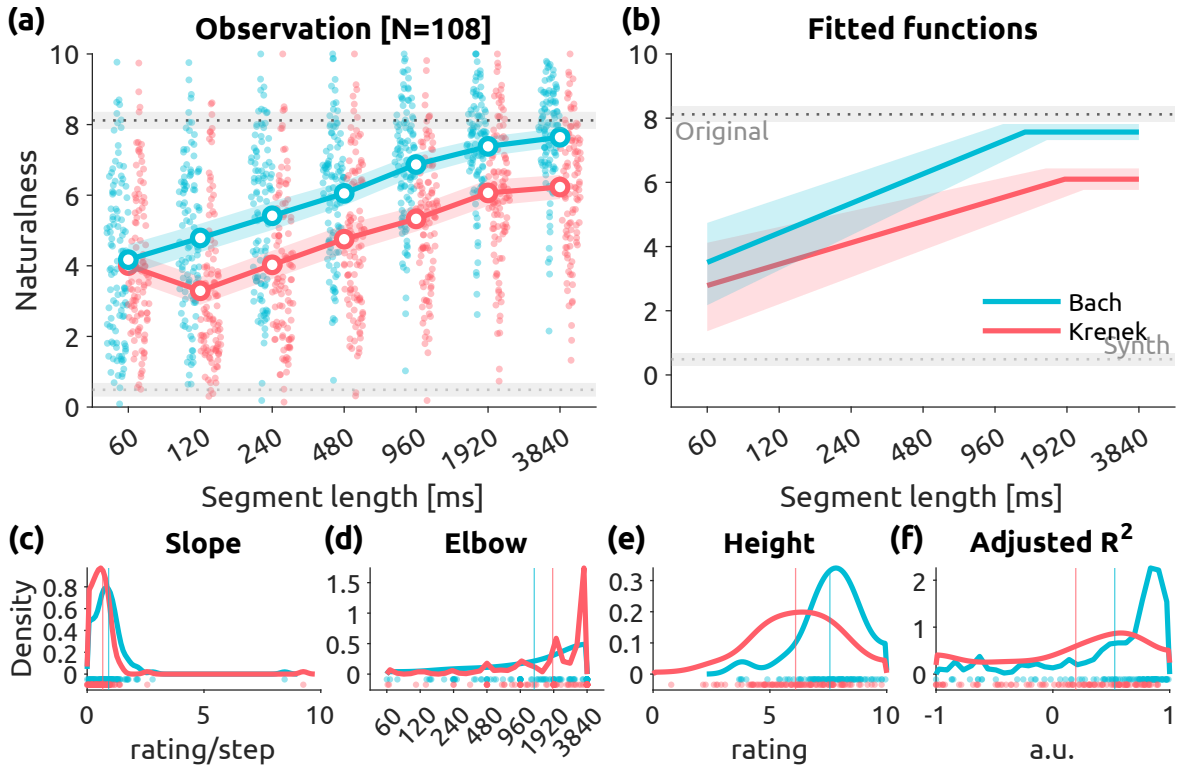

Figure 3: Experiment III: observed means and distribution of fitted parameters. Data are displayed in the identical scheme as in Figure 1.

observed (a steeper slope for Bach, $P_{\mathrm{FDR}} = 0.037$; an elbow approximately half a step lower for Bach, corresponding to ∼500 ms, $P_{\mathrm{FDR}} = 0.001$).

4.2 Discussion

Experiment III again showed a strong main effect of tonality, consistent with the previous two experiments. Interestingly, an initial overshoot in naturalness ratings for the atonal (Krenek) quilts (at 60 ms) was observed, resulting in differences in the fitted elbow function parameters. Specifically, naturalness ratings for atonal quilts saturated earlier than for tonal quilts, and the rate of increase over segment length was higher for tonal than atonal quilts, suggesting a modest interaction between tonality and temporal structure processing, albeit of much smaller magnitude than the main effect of tonality.

The differences between the exemplar sets used in Experiments I and II versus Experiment III were minimal (< 0.58 standard deviations across metrics) in terms of tonal–atonal matching (correlation coefficients of auditory model statistics), quilting quality (mean absolute difference [MAD] of segment transitions),

Table 3: Experiment III: fitted parameters of the elbow function ($N = 108$).

Contrast	Parameter	Mean	Std. Dev.	95% CI	t[107]	P_{FDR}
Bach vs. null	Slope [rating/step]	0.919	0.892	$[0.749, 1.090]$	10.709	$< 10^{-17}$
	Elbow [step]	5.417	1.587	$[5.115, 5.720]$	35.477	$< 10^{-59}$
	Height [rating]	7.568	1.314	$[7.317, 7.818]$	59.846	$< 10^{-82}$
	Adjusted R^2	0.530	0.506	$[0.433, 0.626]$	10.886	$< 10^{-18}$
Krenek vs. null	Slope [rating/step]	0.667	0.916	$[0.492, 0.842]$	7.562	$< 10^{-10}$
	Elbow [step]	5.970	1.359	$[5.710, 6.229]$	45.646	$< 10^{-70}$
	Height [rating]	6.099	1.757	$[5.764, 6.434]$	36.066	$< 10^{-60}$
	Adjusted R^2	0.196	0.555	$[0.090, 0.302]$	3.674	$< 10^{-3}$
Bach vs. Krenek	ΔSlope [rating/step]	0.253	1.240	$[0.016, 0.489]$	2.117	0.037
	ΔElbow [step]	-0.552	1.728	$[-0.882, -0.222]$	-3.321	0.001
	ΔHeight [rating]	1.468	1.182	$[1.243, 1.694]$	12.905	$< 10^{-22}$

and loudness in sones. However, when using different exemplars, even if they were well matched, slight perturbations in estimates are expected, which can be larger when a small set of exemplars is used, as in the present study (K=4). Nonetheless, Experiment III successfully replicated the main effect of tonality.

5 Experiment IV: Cross-domain

Across three experiments with music quilts, we consistently found main effects of segment length and tonality, alongside substantial individual differences in estimated parameters. One possible explanation for these individual differences is that low-level auditory processing mechanisms (Norman-Haignere, Long, et al., 2022b; Stevenson et al., 2012) play a significant role, potentially influenced by genetic factors (Gingras et al., 2015) that affect multiple auditory domains beyond music, including speech perception. Alternatively, individual differences may reflect domain-specific processes in tonal music perception (Peretz & Hyde, 2003), consistent with evidence for music-selective neural populations (Boebinger et al., 2021; Norman-Haignere, Feather, et al., 2022; Norman-Haignere et al., 2015).

To investigate whether the observed effects are domain-general or domain-specific, we invited participants to a new experiment using speech quilts. To maintain the parallelism with music quilts, the speech quilts were created from recordings of English-Korean bilingual speakers, following procedures from our previous work (Kim et al., 2024; Overath & Paik, 2021; Overath et al., 2015). By exclusively recruiting participants with no knowledge of Korean, we were able to compare the effects of temporal degradation on naturalness ratings for familiar (English) versus unfamiliar (Korean) speech, in a manner analogous to the comparison of tonal (Bach) versus atonal (Krenek) music.

5.1 Materials and Methods

5.1.1 Speech Stimuli

Detailed procedures for speech quilt generation are described in the previous work (Kim et al., 2024). Briefly, 8-s speech quilts were generated from recordings of four female bilingual (Korean and English) speakers reading textbook passages. Recordings were first segmented into 60-s excerpts and fed into the same quilting algorithm used for the music stimuli, with global minimisation of artefacts across all possible initialisations. Based on Overath et al. (2015), segment lengths were shorter than for music: 30, 60, 120, 240, 480, and 960 ms. Using an algorithm that matches the first four statistical moments of spectrograms and 2-D decomposed spectrograms (Norman-Haignere & McDermott, 2018), the two best-matching pairs of English and Korean quilts were identified for each speaker (8 exemplars per segment length). Stimuli were loudness-normalised to ∼31 sones (Supplementary Methods S1.4) with 20-ms cosine ramps applied. Detailed metrics are provided in the Supplementary File.

5.1.2 Participant Recruitment

Only MTurk participants who had already taken part in either Experiment I or II and had no knowledge of Korean were re-invited in August 2021.

5.1.3 Procedure

The procedure was similar to Experiments I–III, except that the main experiment comprised 8 blocks of 14 trials each (6 English quilts, 6 Korean quilts, 1 English original, 1 Korean synth). A short questionnaire on region of origin and language proficiency (Korean, Mandarin, Cantonese, Japanese, English) was administered in place of the Gold-MSI questionnaire, as all participants had already completed it in their prior session. Sessions lasted ∼25 min on average. All completed responses were manually approved and compensated at US$6.00, reflecting the higher compensation offered to encourage return participation.

5.1.4 Data Analysis

Of 120 participants, 20 were excluded during post-screening and 1 was excluded for not meeting the language requirement. The remaining 99 participants were included in the subsequent analysis.

Table 4: Experiment IV: fitted parameters of the elbow function ($N = 99$).

Contrast	Parameter	Mean	Std. Dev.	95% CI	t[98]	P_{FDR}
Eng vs. null	Slope [rating/step]	1.413	0.398	$[1.334, 1.493]$	35.309	$< 10^{-56}$
	Elbow [step]	5.822	0.377	$[5.746, 5.897]$	153.536	$< 10^{-116}$
	Height [rating]	6.688	1.838	$[6.321, 7.055]$	36.198	$< 10^{-57}$
	Adjusted R^2	0.837	0.152	$[0.807, 0.867]$	54.934	$< 10^{-74}$
Kor vs. null	Slope [rating/step]	1.668	0.504	$[1.568, 1.769]$	32.953	$< 10^{-53}$
	Elbow [step]	5.524	0.539	$[5.417, 5.632]$	102.010	$< 10^{-99}$
	Height [rating]	7.389	1.907	$[7.009, 7.770]$	38.558	$< 10^{-59}$
	Adjusted R^2	0.858	0.215	$[0.815, 0.901]$	39.602	$< 10^{-60}$
Eng vs. Kor	ΔSlope [rating/step]	-0.255	0.458	$[-0.346, -0.164]$	-5.536	$< 10^{-6}$
	ΔElbow [step]	0.297	0.557	$[0.186, 0.408]$	5.312	$< 10^{-6}$
	ΔHeight [rating]	-0.701	1.864	$[-1.073, -0.330]$	-3.745	$< 10^{-3}$

5.2 Results

5.2.1 Naturalness Ratings Increased over Segment Lengths, Interacting with Speech Familiarity

Table 4 and Figure 4 show the descriptive and inferential statistics for naturalness ratings of speech quilts. Clear main effects of segment length and familiarity, and an interaction between them, were found ($P_{\mathrm{FDR}} < 10^{-3}$). However, the directions of these effects were opposite to those in Experiments I–III, if we treat tonal music as analogous to the native language and atonal music as analogous to a foreign language. Specifically, the plateau height was lower for English than Korean ($P_{\mathrm{FDR}} < 10^{-3}$), and the slope was steeper and the elbow point lower for Korean ($P_{\mathrm{FDR}} < 10^{-6}$)—the reverse of the pattern observed in Experiment I–III.

5.2.2 Correlation Between Familiarity Effects in Music and Speech

We investigated whether the observed effects of temporal degradation and their interaction with familiarity correlate across domains. To this end, we calculated the means and differences (familiar-minus-unfamiliar) of estimated parameters within each participant and correlated these across domains. With one outlier removed for extreme values ($|Z| > 5$), the mean elbow showed a weak correlation ($r[92] = 0.225$, uncorrected $P = 0.025$, Figure 5). However, no parameter survived the multiple comparison correction, even when controlling for the age group ($\min\ P_{\mathrm{FDR}} = 0.082$; Table 5). Given the ceiling effect in speech elbow estimates (Figure 5b), Spearman's rank correlation was additionally computed, but no correlation was found ($\min\ P_{\mathrm{FDR}} = 0.128$).

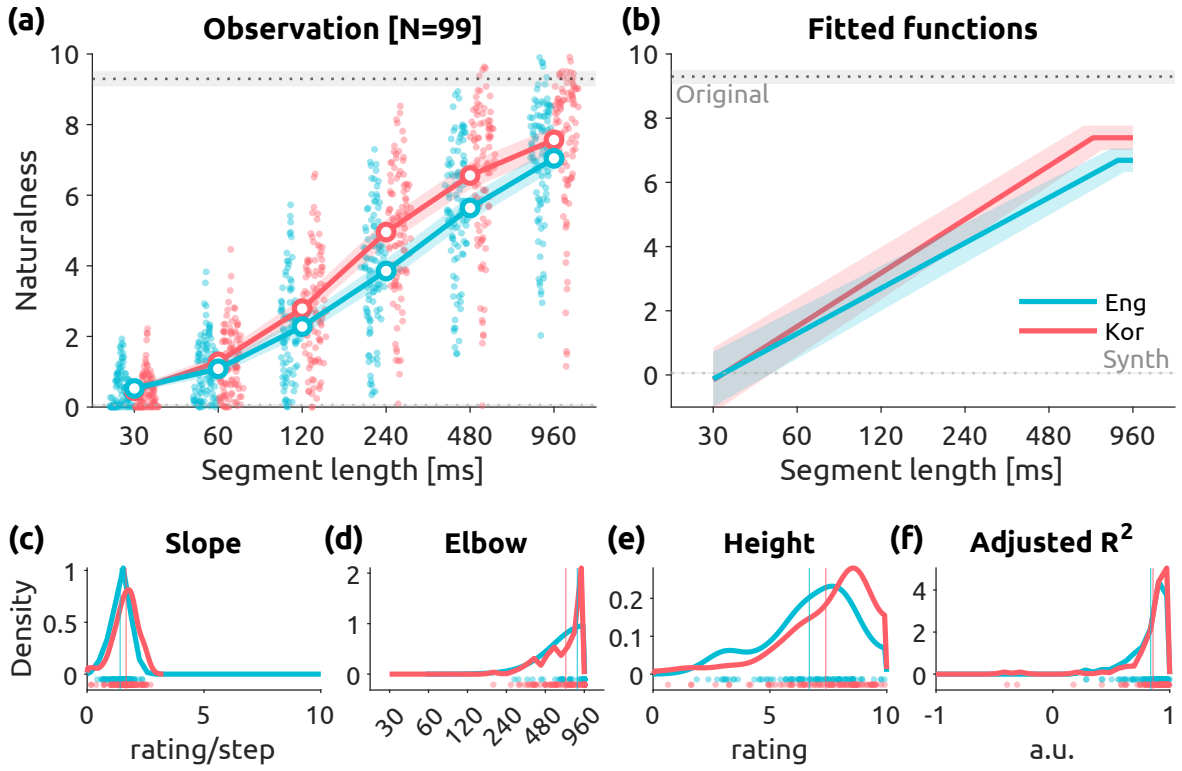

Figure 4: Experiment IV: observed means and distribution of fitted parameters. Data are displayed in the identical scheme as in Figure 1 except for the types of stimuli: English [blue] and Korean [red].

Table 5: Experiment IV: inferential statistics of correlations between music and speech domains in non-musicians ($N = 99$). Age represents a 4-level categorical variable of age group.

Model	Estimate	95%CI	t[93]	P_{FDR}
Slope(music) $\sim 1 +$ Slope(speech) $+$ Age	0.110	$[-0.129, 0.348]$	0.915	0.36
Elbow(music) $\sim 1 +$ Elbow(speech) $+$ Age	0.598	$[0.077, 1.119]$	2.279	0.08
Height(music) $\sim 1 +$ Height(speech) $+$ Age	0.091	$[-0.065, 0.246]$	1.157	0.30
ΔSlope(music) $\sim 1 + \Delta$Slope(speech) $+$ Age	-0.349	$[-0.657, -0.040]$	-2.244	0.08
ΔElbow(music) $\sim 1 + \Delta$Elbow(speech) $+$ Age	0.459	$[-0.153, 1.072]$	1.489	0.21
ΔHeight(music) $\sim 1 + \Delta$Height(speech) $+$ Age	-0.129	$[-0.280, 0.023]$	-1.690	0.19

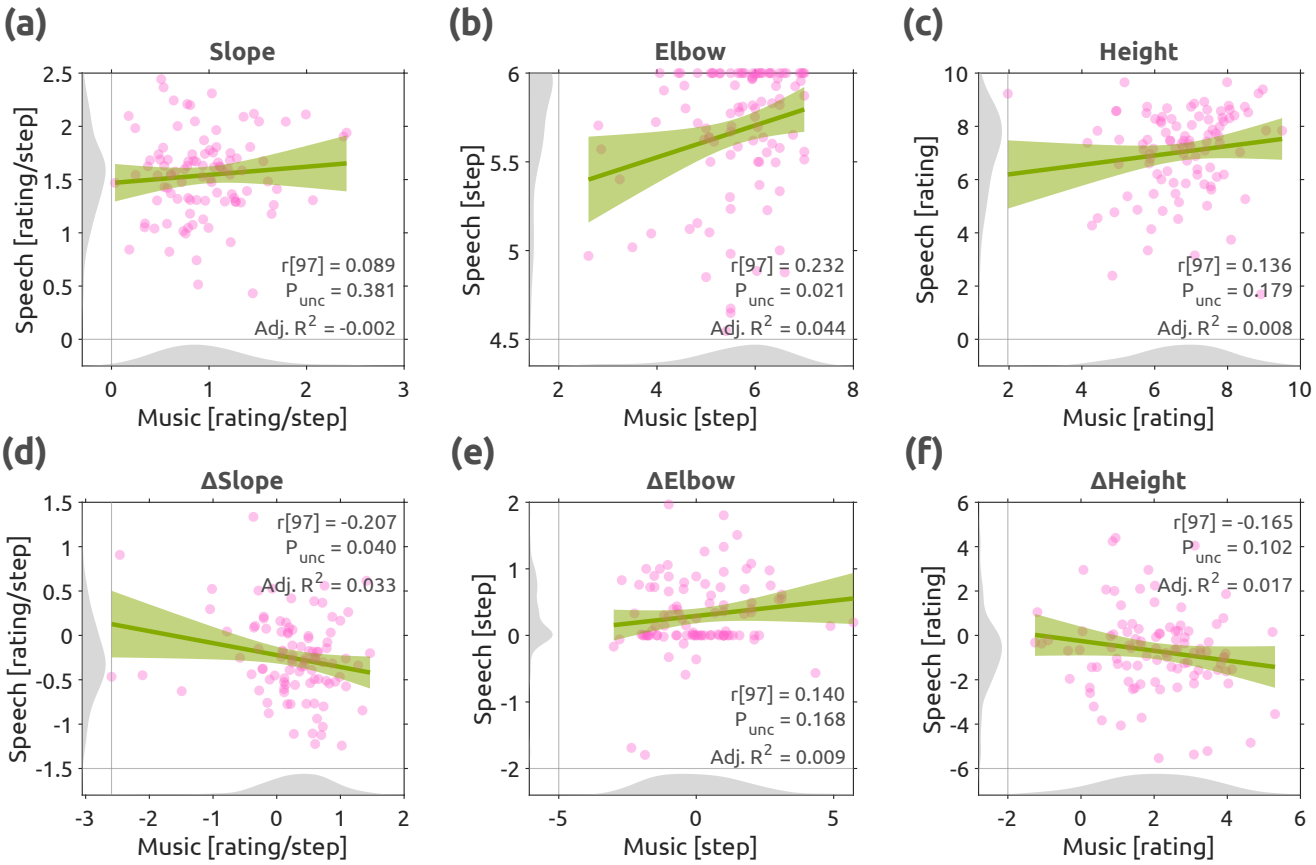

Figure 5: Experiment IV: correlation between music and speech domains in non-musicians ($N = 99$). For each elbow-function's means [(a) slope, (b) elbow, (c) height] or differences [(d) slope, (e) elbow, (f) height] between tonal (or familiar) and atonal (or unfamiliar) quilts, metrics in speech are plotted over metrics in music. Fitted linear functions with 95% confidence intervals of parameter are shown in green. Individuals are shown as magenta dots. Marginal distributions of the differences are shown in gray density plots along the axes.

5.3 Discussion

Consistent with the music quilts, naturalness ratings of speech quilts increased over segment lengths, indicating that listeners were sensitive to temporal degradation in both domains. However, the reversed effect of familiarity in speech—where English quilts were rated as less natural than Korean quilts—contrasts with the higher naturalness ratings for tonal (Bach) than atonal (Krenek) music. A possible explanation is that native English speakers are more sensitive to disruptions in the temporal structure of their native language than to disruptions in tonal music, particularly given the below-average musical sophistication reported by most participants. Whilst disruptions in 960-ms English quilts were perceptually salient, subtle disruptions in Bach quilts may have been less detectable for this population.

Regarding the cross-domain correlation, we found no strong evidence for a shared mechanism of temporal structure perception across music and speech. Whilst reversed, this result may suggest domain-specificity in processing short (60–3840 ms) temporal contexts in music. It also remains possible that ecologically salient sounds such as voice are processed differently from other complex sounds. Taken together, the current findings suggest that naturalness ratings may have reflected different perceptual dimensions depending on the type of information available: linguistic knowledge in speech, and tonal knowledge in music.

6 Experiment V: Musicianship

In Experiment IV, we observed a reversed effect of familiarity in speech relative to music. One potential explanation is heightened sensitivity in language perception, driven by higher-level linguistic knowledge (e.g., semantics and syntax), compared to music perception in non-musicians. To probe this hypothesis, we conducted an experiment with musicians who possessed greater familiarity with Western tonal music. Under the familiarity hypothesis, musicians with extensive training in Western tonal music were expected to show heightened sensitivity to temporal degradation in music, producing a tonality effect more akin to the familiarity effect observed in speech.

6.1 Materials and Methods

6.1.1 Participant Recruitment

Musicians were recruited via advertisements distributed to local music communities (e.g., orchestras) in the Durham, North Carolina area between December 2021 and March 2023. Inclusion criteria were: (a) native English speaker with no knowledge of Korean, and (b) high familiarity with classical music. Musical background was verified by a questionnaire (see Procedure below).

6.1.2 Procedure

Participants completed two sessions from Experiment I (music quilts) and Experiment IV (speech quilts). The order of sessions was counterbalanced across participants. Questionnaires from Experiments I and IV were completed after the sessions. The whole procedure took $\sim$1 hour. Participants were compensated with a US$10 Amazon gift card.

6.1.3 Analysis

Of 27 participants, 7 were excluded during post-screening and 1 was excluded for not meeting the language requirement. The remaining 19 participants were included in the subsequent analysis. In addition to the analyses used in previous experiments, linear models were fitted to test the main effect of musicianship on estimated parameters and its interactions with tonality or familiarity, whilst covarying age group and experiment cohort where applicable.

6.2 Results

6.2.1 Sample Demographics

Nineteen participating musicians were mostly (84%) between 18 and 29 years old, younger than the MTurk participants (predominantly 30–44 years old). The racial distribution also differed, with under-representation of African ancestry in the musician sample. Gender distribution was comparable across groups.

As for musical sophistication, reported Active Engagement and Perceptual Abilities did not differ from UK population norms ($P > 0.292$), whilst reported Musical Training was significantly higher ($t[18.03] = 3.87$,

Table 6: Experiment V-Music: fitted parameters of the elbow function in musicians ($N = 19$).

Contrast	Parameter	Mean	Std. Dev.	95% CI	t[18]	P_{FDR}
Bach vs. null	Slope [rating/step]	1.210	0.651	$[0.929, 1.492]$	8.918	$< 10^{-7}$
	Elbow [step]	5.839	1.188	$[5.325, 6.353]$	23.563	$< 10^{-15}$
	Height [rating]	7.776	1.784	$[7.005, 8.548]$	20.900	$< 10^{-14}$
	Adjusted R^2	0.631	0.442	$[0.440, 0.822]$	6.843	$< 10^{-5}$
Krenek vs. null	Slope [rating/step]	1.042	0.788	$[0.701, 1.383]$	6.342	$< 10^{-5}$
	Elbow [step]	5.278	2.079	$[4.378, 6.177]$	12.173	$< 10^{-10}$
	Height [rating]	6.241	2.149	$[5.312, 7.170]$	13.928	$< 10^{-11}$
	Adjusted R^2	0.536	0.487	$[0.325, 0.746]$	5.276	$< 10^{-4}$
Bach vs. Krenek	ΔSlope [rating/step]	0.168	0.561	$[-0.075, 0.411]$	1.436	0.165
	ΔElbow [step]	0.562	1.554	$[-0.110, 1.234]$	1.733	0.107
	ΔHeight [rating]	1.535	1.505	$[0.884, 2.186]$	4.892	$< 10^{-4}$

Table 7: Experiment V-Speech: fitted parameters of the elbow function in musicians ($N = 19$).

Contrast	Parameter	Mean	Std. Dev.	95% CI	t[18]	P_{FDR}
Eng vs. null	Slope [rating/step]	1.347	0.604	$[1.056, 1.639]$	9.718	$< 10^{-7}$
	Elbow [step]	5.883	1.290	$[5.261, 6.505]$	19.874	$< 10^{-12}$
	Height [rating]	8.207	1.035	$[7.708, 8.706]$	34.553	$< 10^{-16}$
	Adjusted R^2	0.725	0.262	$[0.599, 0.851]$	12.078	$< 10^{-8}$
Kor vs. null	Slope [rating/step]	1.310	1.150	$[0.756, 1.864]$	4.966	$< 10^{-3}$
	Elbow [step]	5.478	1.930	$[4.548, 6.408]$	12.369	$< 10^{-9}$
	Height [rating]	6.653	1.748	$[5.811, 7.496]$	16.587	$< 10^{-11}$
	Adjusted R^2	0.599	0.381	$[0.415, 0.782]$	6.858	$< 10^{-5}$
Eng vs. Kor	ΔSlope [rating/step]	0.038	1.079	$[-0.482, 0.557]$	0.152	0.881
	ΔElbow [step]	0.405	1.395	$[-0.267, 1.077]$	1.266	0.244
	ΔHeight [rating]	1.554	1.583	$[0.791, 2.317]$	4.279	$< 10^{-3}$

$P = 0.001$). Musicians reported higher musical sophistication than participants in previous experiments across all subscales (see Table S2). Two musicians reported growing up in the Asian region; all others reported growing up in North America.

6.2.2 Naturalness Ratings Increased over Segment Lengths, Interacting with Familiarity

Similarly to Experiments I–III, the main effect of segment length was clear ($P_{\mathrm{FDR}} < 10^{-3}$; Table 6, Figure 6). When comparing Bach and Krenek quilts, plateau heights differed, with Bach higher than Krenek ($P_{\mathrm{FDR}} < 10^{-3}$), consistent with the results from non-musicians.

Similarly to Experiment IV, a marked main effect of segment length was found ($P_{\mathrm{FDR}} < 10^{-3}$) and the higher plateau height in Korean than English ($P_{\mathrm{FDR}} < 10^{-3}$; Table 7 and Figure 7).

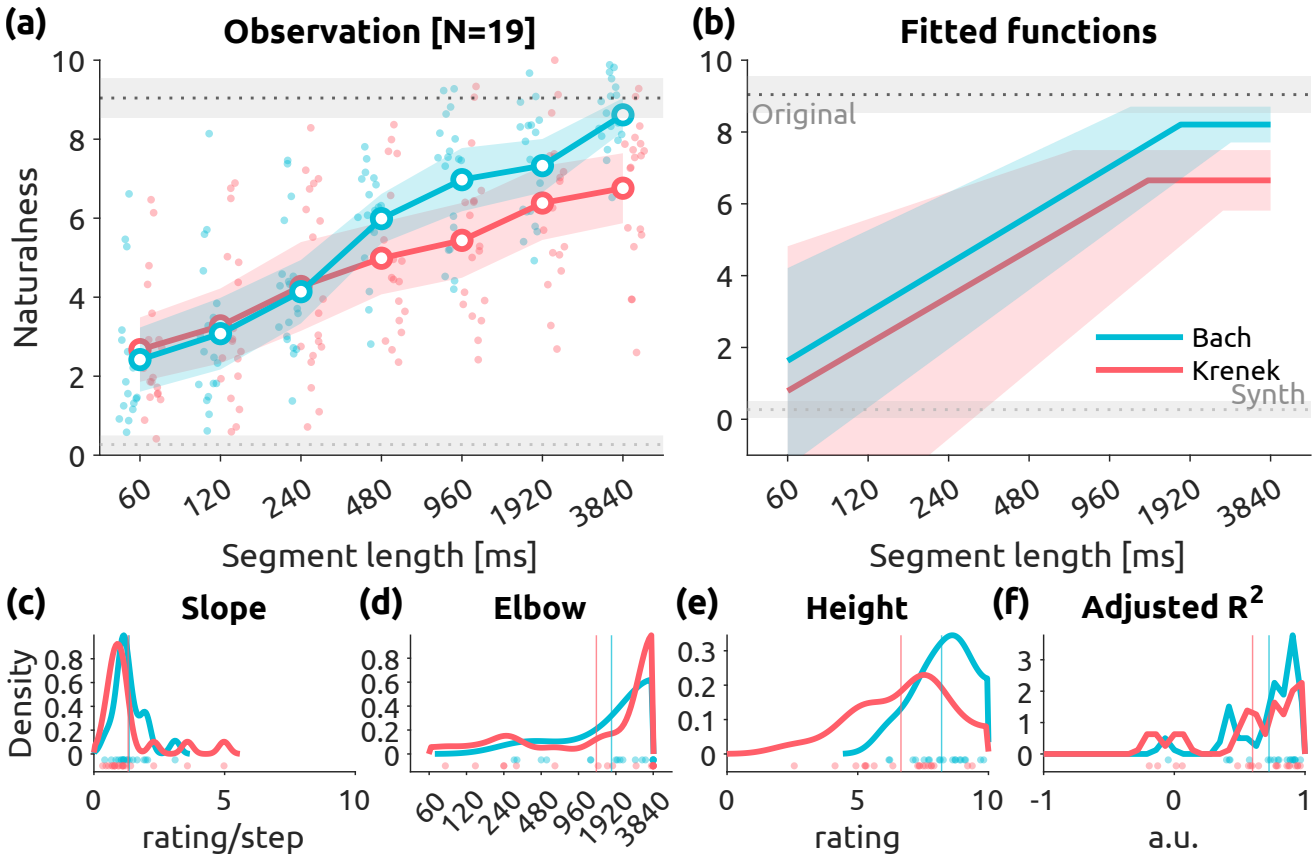

Figure 6: Experiment V-Music: observed means and distribution of fitted parameters. Data are displayed in the identical scheme as in Figure 1.

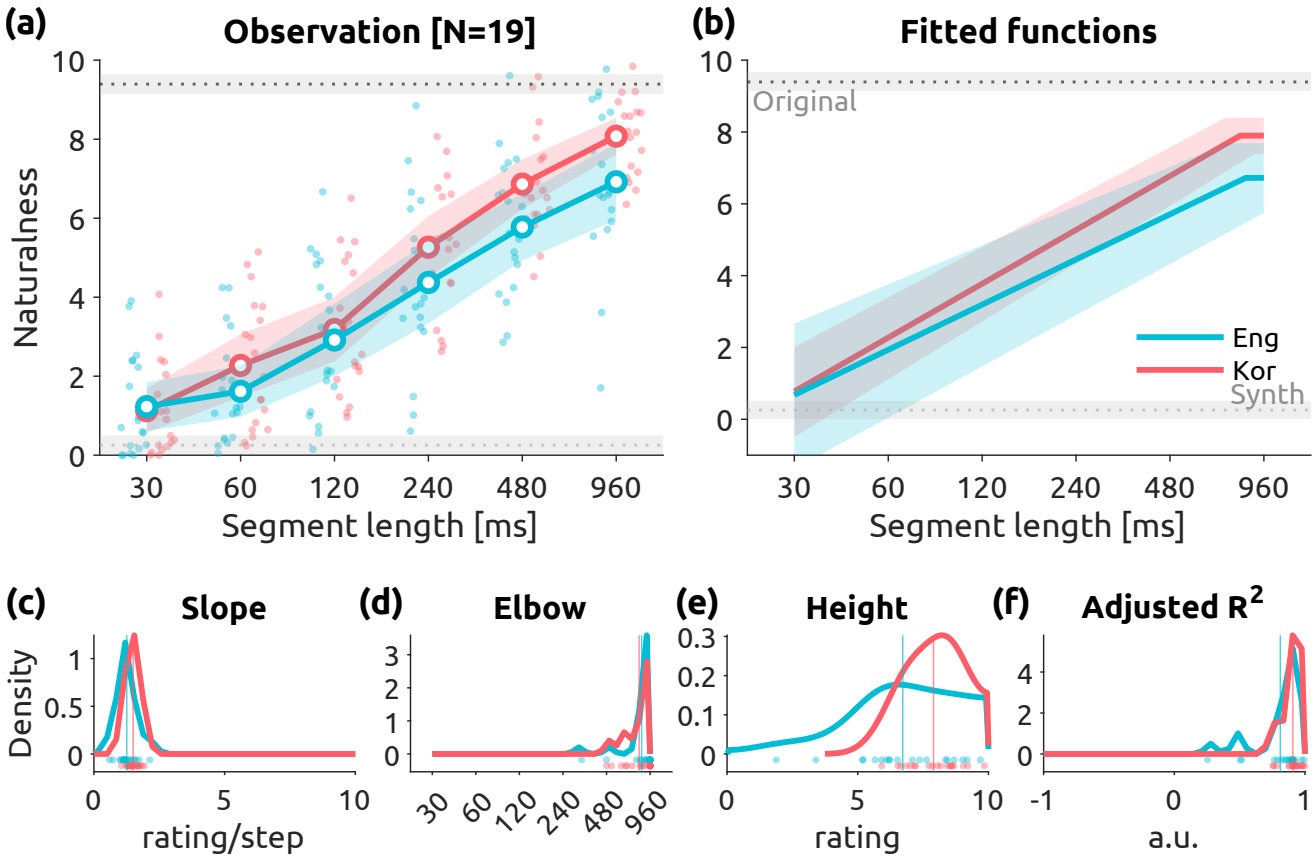

Figure 7: Experiment V-Speech: observed means and distribution of fitted parameters. Data are displayed in the identical scheme as in Figure 4.

Table 8: Experiment V: inferential statistics of correlations between music and speech domains in musicians ($N = 19$). Age represents a 4-level categorical variable of age group.

Model	Estimate	95%CI	t[93]	P_{FDR}
Slope(music) $\sim$ 1 + Slope(speech) + Age	-0.664	$[-2.669, 1.341]$	-0.702	0.74
Elbow(music) $\sim$ 1 + Elbow(speech) + Age	-0.065	$[-2.810, 2.680]$	-0.050	0.96
Height(music) $\sim$ 1 + Height(speech) + Age	0.264	$[-0.243, 0.772]$	1.104	0.74
ΔSlope(music) $\sim$ 1 + ΔSlope(speech) + Age	-0.484	$[-1.762, 0.793]$	-0.804	0.74
ΔElbow(music) $\sim$ 1 + ΔElbow(speech) + Age	-0.763	$[-2.417, 0.892]$	-0.977	0.74
ΔHeight(music) $\sim$ 1 + ΔHeight(speech) + Age	-0.094	$[-0.509, 0.320]$	-0.481	0.77

6.2.3 Correlation Between Music and Speech

Following the same analysis scheme for investigating individual-level associations between responses to music and speech stimuli as described in Section 5.2.2, no strong correlation between music and speech parameters was found after multiple comparison correction, accounting for the age group ($\min\ P_{\mathrm{FDR}} = 0.740$; Table 8, Figure 8).

6.2.4 Comparison Between Musicians and Non-musicians

First, we confirmed that there were no mean-level differences in ratings between groups after accounting for age group in a linear model (musicians vs. non-musicians; music quilts: $t[318] = 0.941$, $P = 0.347$; speech quilts: $t[112] = 1.344$, $P = 0.182$). Subsequently, estimated effects were compared between musicians and non-musicians using ordinary least squares (OLS) after removing six non-musician outliers with noisy estimates ($|Z| > 5$). Among other effects, the slopes and heights for Krenek were significantly higher in musicians than non-musicians ($P_{\mathrm{FDR}} < 0.05$; Table 9). No interaction between musicianship and the effect of tonality (or familiarity) was found ($\min\ P_{\mathrm{FDR}} = 0.199$; Table 10).

However, Figure 9 clearly shows a larger sampling variance in musicians due to the small sample size, suggesting that the absence of a group difference is not strongly supported.

6.3 Discussion

Contrary to the familiarity hypothesis, which predicted that sensitivity to temporal degradation would be modulated by familiarity with tonal music, musicians reporting markedly higher musical sophistication showed a similar effect of tonality to non-musicians (i.e., a higher plateau for Bach than Krenek). Whilst appropriate caution is warranted given the relatively small musician sample, the effect of familiarity in speech and music perception may nonetheless differ. In particular, the semantic and syntactic structure

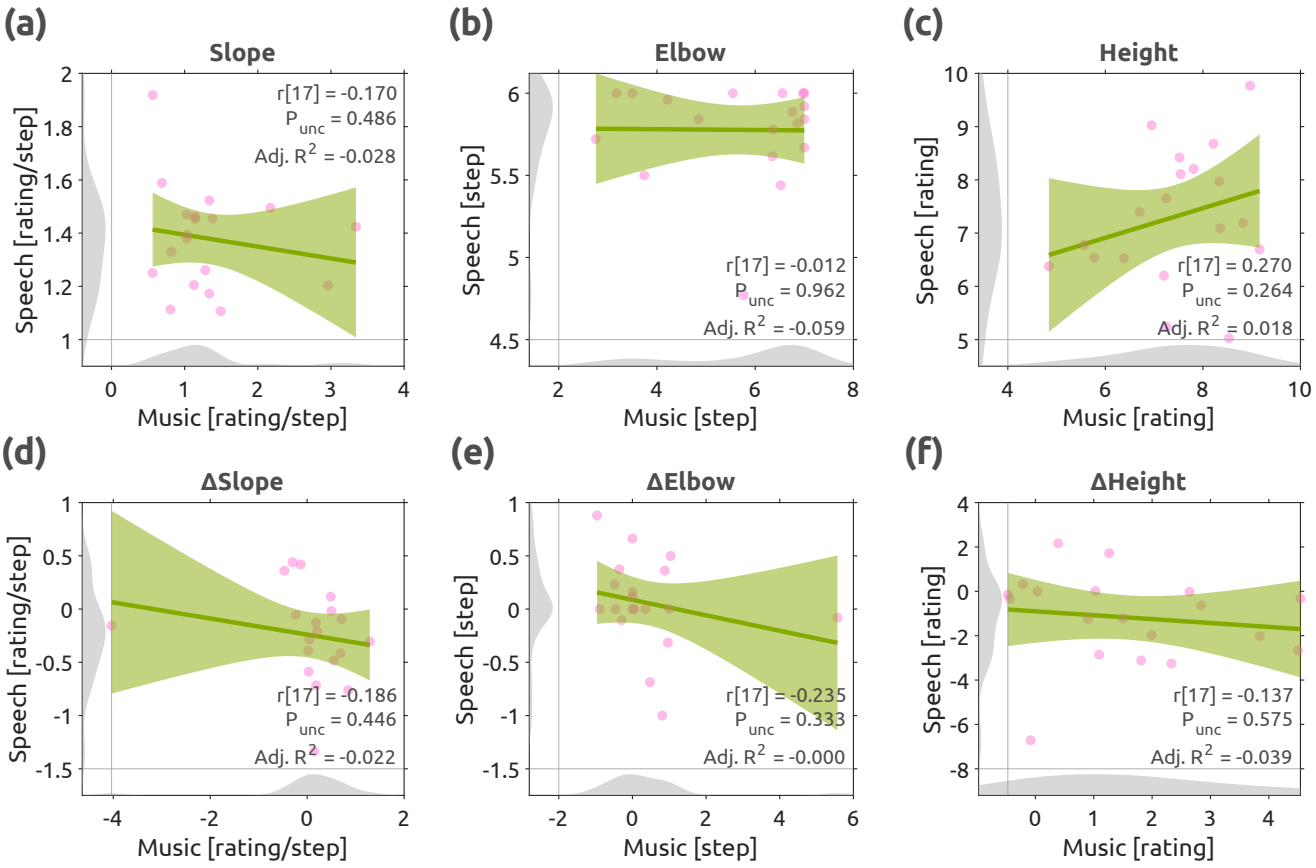

Figure 8: Experiment V: correlation between music and speech domains in musicians ($N = 19$). Data are displayed in the identical scheme as in Figure 5.

Table 9: Comparison of parameters between musicians and nonmusicians. Parameter names, differences in parameters (Musician-minus-Nonmusician), 95% confidence interval (CI), t-statistic, degrees of freedom (df), and FDR-adjusted P-values are displayed.

Metric	Diff (Mus-NMus)	95% CI	t-stat.	df	P_{FDR}
Slope (Bach)	0.241	$[-0.037, 0.519]$	1.705	312.000	0.256
Elbow (Bach)	0.291	$[-0.374, 0.956]$	0.861	312.000	0.546
Height (Bach)	0.642	$[0.055, 1.230]$	2.152	312.000	0.129
Slope (Krenek)	0.574	$[0.242, 0.906]$	3.403	312.000	0.009
Elbow (Krenek)	−0.328	$[-1.085, 0.428]$	−0.854	312.000	0.546
Height (Krenek)	1.199	$[0.393, 2.005]$	2.926	312.000	0.022
Slope (English)	−0.079	$[-0.295, 0.136]$	−0.729	112.000	0.546
Elbow (English)	−0.072	$[-0.284, 0.140]$	−0.675	112.000	0.546
Height (English)	0.152	$[-0.887, 1.190]$	0.289	112.000	0.773
Slope (Korean)	−0.106	$[-0.371, 0.158]$	−0.795	112.000	0.546
Elbow (Korean)	0.235	$[-0.051, 0.520]$	1.627	112.000	0.256
Height (Korean)	0.559	$[-0.451, 1.569]$	1.096	112.000	0.546

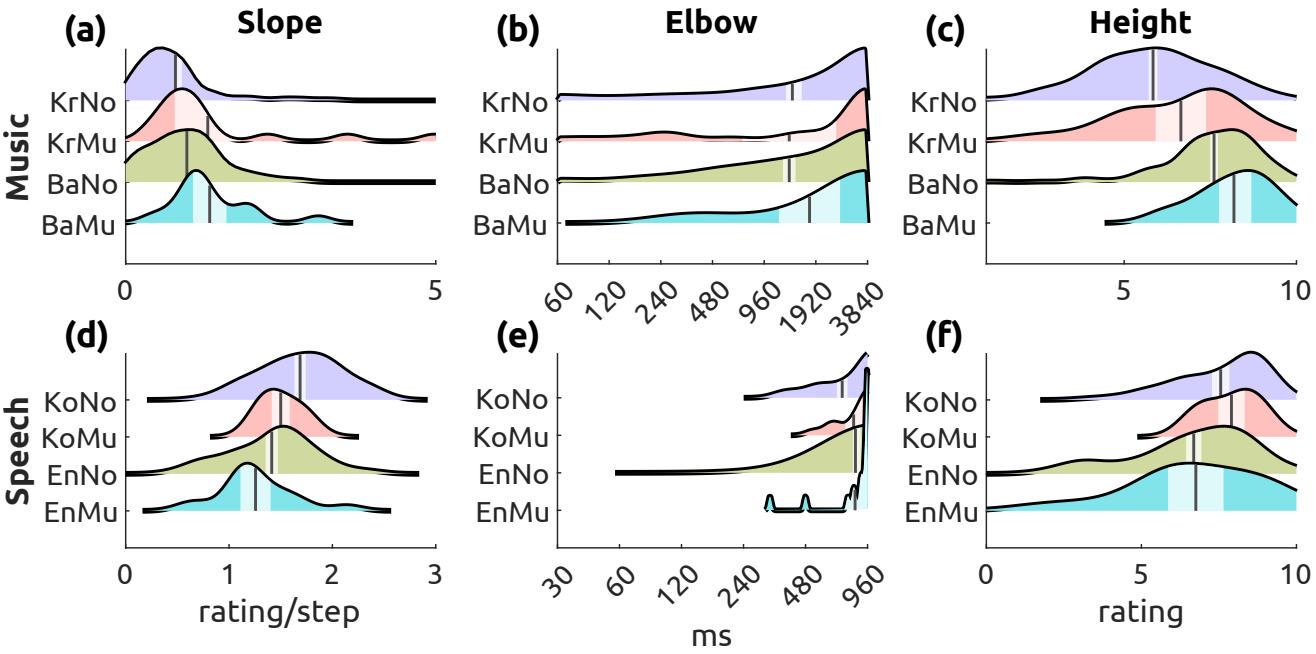

Figure 9: Ridgeline plots display parameter densities of musicians and nonmusicians. White vertical bands mark 95% confidence intervals and grey vertical lines mark group means. Abbreviations: BaMu: Bach-Musicians, BaNo: Bach-Nonmusicians, KrMu: Krenek-Musicians, KrNo: Krenek-Nonmusicians, EnMu: English-Musicians, EnNo: English-Nonmusicians, KoMu: Korean-Musicians, KoNo: Korean-Nonmusicians.

Table 10: Interaction between tonality/familiarity and musicianship. Names of contrasts, parameters, sum of squares (SS), partial eta squared (η_p^2), F-statistic, degrees of freedom (df), and FDR-adjusted P-value are displayed.

Contrast	Metric	SS	η_p^2	F-stat	df	P_{FDR}
Tonality:Musicianship	Slope (Music)	0.905	0.009	2.870	$[1, 312]$	0.199
	Elbow (Music)	3.128	0.006	1.777	$[1, 312]$	0.275
	Height (Music)	2.522	0.009	2.733	$[1, 312]$	0.199
Familiarity:Musicianship	Slope (Speech)	0.005	0.001	0.044	$[1, 112]$	0.834
	Elbow (Speech)	0.599	0.035	4.096	$[1, 112]$	0.199
	Height (Speech)	1.056	0.005	0.598	$[1, 112]$	0.529

available in the native language may have contributed to naturalness ratings in speech in ways that have no direct analogue in music perception.

Rather than the expected heightened sensitivity to tonal music (i.e., lower naturalness ratings for Bach quilts), musicians showed elevated slopes and plateau heights for Krenek quilts relative to non-musicians. This suggests that musicians perceived atonal music as more natural when temporal structure was largely intact (higher plateau) and responded more sensitively to its degradation (higher slope).

7 Association with Self-reported Musicality

Although we did not find associations in estimated parameters between music and speech, individual differences in the perception of the temporal or tonal structure of music and speech may nonetheless be associated with different facets of musical sophistication as measured by self-report. Here, we further explore associations between estimated parameters from the perceptual curves of participants and their Gold-MSI subscale scores by aggregating responses from non-musicians, and demonstrate the replicability of the found associations using cross-validation.

7.1 Materials and Methods

7.1.1 Participant Recruitment

For music quilts, participants from Experiments I–III ($N = 304$) were aggregated to investigate associations between fitted parameters and demographic and self-reported musicality variables. For speech quilts, participants from Experiment IV ($N = 98$) were used.

7.1.2 Scales

Gold-MSI is a self-report scale of musical sophistication designed for the general population (Müllensiefen et al., 2014). The overall score has shown good convergent validity with behavioural measures of musical aptitude (Gordon's Advanced Measures of Audiation; $r = 0.33$–0.51; Müllensiefen et al., 2014), an online behavioural test (Musical Ear Test; $r = 0.47$; Correia et al., 2022), a working memory precision task for frequency ($r = 0.50$; Lad et al., 2022), and an auditory-motor coupling task ($r = 0.26$; Rimmele et al., 2022). In the current study, we used two subscales (Perceptual Abilities and Musical Training) that are highly predictive of melodic memory and beat perception, and one subscale (Active Engagement) as an

index of potential musical exposure or informal and implicit musical skill acquisition (Müllensiefen et al., 2014).

7.1.3 Data Analysis

A linear model controlling for age group and experiment cohort was fit for music quilts for each of the three Gold-MSI subscales as dependent variable:

$$\mathrm{GMSI} \sim 1 + \mathrm{Param} + \mathrm{Age} + \mathrm{Exp} \tag{2}$$

where GMSI is one of three Gold-MSI subscales (response variable), Param is one of nine fitted parameters ([Bach, Krenek, Bach-minus-Krenek] $\times$ [Slope, Elbow, Height]) (predictor variable), Age is a 4-level factor (18–29, 30–44, 45–59, 60+), and Exp is a 3-level factor (Exp I, II, III) for the experiment number. Age and experiment cohort were included as covariates given significant associations (Supplementary Results S 2.2). To control the family-wise Type I error rate due to multiple testing ($M = 18$), FDR adjustment was applied (Benjamini & Hochberg, 1995).

To assess the reproducibility of observed associations, a repeated 2-fold cross-validation was used. Specifically, a linear model was fitted using ordinary least squares (OLS) on a randomly split half of participants. The estimated coefficients were then applied to predict musicality variables with covariates on the held-out half, evaluating prediction accuracy using mean absolute error (MAE) and Pearson correlation between actual and predicted values. This random-split procedure was iterated 1,000 times, with experiment groups stratified to ensure proportional representation in both training and test sets (Hastie et al., 2009). For each split, a null model (without Param) and the model of interest (with Param) were fitted, and a "winning model" was determined by comparing MAEs.

For the speech quilts, since a single experiment cohort was used, a linear model without the cohort covariate was fit.

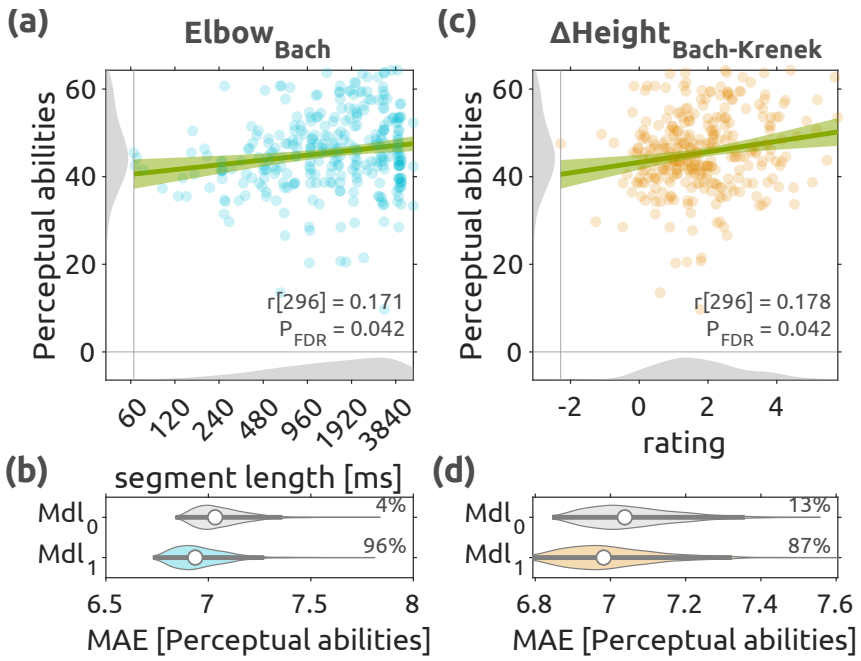

Figure 10: Correlations between fitted parameters and a self-reported musicality ($N = 304$). (a, c) Scatter plots of the Perceptual Abilities scale over elbow function parameters. Variables are adjusted for covariates. (b, d) Violin plots of mean absolute errors (MAE) in predicting the Perceptual Abilities with (Mdl_1) or without (Mdl_0) the parameter.

7.2 Results

7.2.1 Perception of Temporal Structure in Tonal Music Was Associated with Self-reported Perceptual Abilities

We found robust associations of reported Perceptual Abilities with the elbow parameter for Bach quilts ($r[296] = 0.171$, $P_{\text{FDR}} = 0.042$) and with the difference in plateau heights between Bach and Krenek ($r[296] = 0.178$, $P_{\text{FDR}} = 0.042$; Table 11, Figure 10). These associations were successfully replicated in cross-validation (the model of interest won in 96% and 87% of random splits, respectively; Figure 10). No associations were found between any parameters of the perceptual functions and any of the other Gold-MSI subscales (Musical Training or Active Engagement)

For speech quilts, no associations reached significance after FDR adjustment. In particular, associations with Perceptual Abilities were only weak (uncorrected $P \geq 0.035$, $r \leq 0.218$, elbow in Korean; Table 12).

Table 11: Associations between fitted parameters of Music quilts and self-reported musicality metrics with age group and experiment cohort covaried ($N = 304$). The number of binary covariates was 5. Abbreviations: AE, Active Engagement; PA, Perceptual Abilities; MT, Musical Training

Gold-MSI	Metrics	t[297]	r[296]	P_{FDR}	Adj. R^2
AE	Slope (Bach)	-0.102	-0.006	0.959	0.057
	Elbow (Bach)	-0.321	-0.019	0.959	0.057
	Height (Bach)	-0.236	-0.014	0.959	0.057
	Slope (Krenek)	1.820	0.105	0.338	0.067
	Elbow (Krenek)	-1.510	-0.087	0.356	0.064
	Height (Krenek)	-0.056	-0.003	0.959	0.057
	ΔSlope (Bach-Krenek)	-1.585	-0.092	0.356	0.065
	ΔEblow (Bach-Krenek)	1.007	0.058	0.567	0.060
	ΔHeight (Bach-Krenek)	-0.155	-0.009	0.959	0.057
PA	Slope (Bach)	1.745	0.101	0.338	0.016
	Elbow (Bach)	2.982	0.171	0.042	0.035
	Height (Bach)	1.737	0.100	0.338	0.016
	Slope (Krenek)	0.300	0.017	0.959	0.006
	Elbow (Krenek)	1.910	0.110	0.338	0.018
	Height (Krenek)	-1.149	-0.067	0.523	0.010
	ΔSlope (Bach-Krenek)	0.730	0.042	0.786	0.008
	ΔEblow (Bach-Krenek)	0.587	0.034	0.836	0.007
	ΔHeight (Bach-Krenek)	3.118	0.178	0.042	0.037
MT	Slope (Bach)	1.549	0.090	0.356	0.005
	Elbow (Bach)	-1.714	-0.099	0.338	0.007
	Height (Bach)	-1.028	-0.060	0.567	0.001
	Slope (Krenek)	0.622	0.036	0.836	-0.001
	Elbow (Krenek)	-1.459	-0.084	0.358	0.004
	Height (Krenek)	0.144	0.008	0.959	-0.003
	ΔSlope (Bach-Krenek)	0.350	0.020	0.959	-0.002
	ΔEblow (Bach-Krenek)	-0.051	-0.003	0.959	-0.003
	ΔHeight (Bach-Krenek)	-1.156	-0.067	0.523	0.002

Table 12: Associations between fitted parameters of Speech quilts and self-reported musicality metrics with age group covaried ($N = 98$). The number of binary covariates was 3. Abbreviations: AE, Active Engagement; PA, Perceptual Abilities; MT, Musical Training.

Gold-MSI	Metrics	t[93]	r[92]	P_{FDR}	Adj. R^2
AE	Slope (Eng)	-0.037	-0.004	0.970	-0.001
	Elbow (Eng)	-2.169	-0.221	0.300	0.047
	Height (Eng)	-1.216	-0.126	0.513	0.015
	Slope (Kor)	-1.783	-0.183	0.300	0.032
	Elbow (Kor)	1.434	0.148	0.513	0.021
	Height (Kor)	-1.816	-0.186	0.300	0.033
	ΔSlope (Eng-Kor)	1.959	0.200	0.300	0.039
	ΔEblow (Eng-Kor)	-2.918	-0.291	0.119	0.083
	ΔHeight (Eng-Kor)	0.662	0.069	0.765	0.004
PA	Slope (Eng)	-0.206	-0.021	0.920	-0.009
	Elbow (Eng)	0.314	0.033	0.920	-0.009
	Height (Eng)	-1.098	-0.114	0.513	0.003
	Slope (Kor)	0.054	0.006	0.970	-0.010
	Elbow (Kor)	2.141	0.218	0.300	0.038
	Height (Kor)	0.187	0.019	0.920	-0.009
	ΔSlope (Eng-Kor)	-0.238	-0.025	0.920	-0.009
	ΔEblow (Eng-Kor)	-1.858	-0.190	0.300	0.026
	ΔHeight (Eng-Kor)	-1.291	-0.133	0.513	0.008
MT	Slope (Eng)	-0.356	-0.037	0.920	-0.036
	Elbow (Eng)	-1.198	-0.124	0.513	-0.021
	Height (Eng)	-0.837	-0.087	0.643	-0.029
	Slope (Kor)	-1.075	-0.111	0.513	-0.024
	Elbow (Kor)	0.292	0.030	0.920	-0.036
	Height (Kor)	-1.105	-0.114	0.513	-0.024
	ΔSlope (Eng-Kor)	0.887	0.092	0.637	-0.028
	ΔEblow (Eng-Kor)	-1.085	-0.112	0.513	-0.024
	ΔHeight (Eng-Kor)	0.314	0.033	0.920	-0.036

7.3 Discussion

We found robust associations between self-reported Perceptual Abilities and both the saturation point (elbow parameter) for Bach quilts and the Bach-minus-Krenek difference in plateau height. The positive association of Perceptual Abilities with these two parameters might indicate a broader discrimination ability and increased sensitivity to the temporal microstructure for familiar music in individuals with greater musical perception skills. The Perceptual Abilities subscale has shown good predictive validity with behavioural performance, particularly for pitch-related tasks (Correia et al., 2022; Müllensiefen et al., 2014). Our finding of an association with temporal structure perception in tonal music is consistent with this literature. Moreover, the specificity of this association to music but not speech suggests domain-specific temporal processing. However, the speech quilt sample was only one third the size of the music quilt sample. Given previously reported associations between general Gold-MSI scores and a speech production task (Rimmele et al., 2022), the possibility of associations with temporal structure perception in speech warrants further investigation.

8 General Discussion

Temporal structures are fundamental building blocks of auditory communicative systems in both music and speech. In particular, temporal and tonal structures in music have been proposed as crucial factors in inducing aesthetic experience (Kim et al., 2019; Lalitte et al., 2009; Tillmann & Bigand, 1996). The current study investigated how temporal structures in tonal music are perceived. Across a series of experiments, we used atonal music as a musically matched control and employed a highly sophisticated scrambling algorithm to precisely control the amount of intact temporal information in violin solo music. Exploiting the scalability of online experiments, we were able to address these questions using large-scale behavioural data. Figure 11 and Figure 12 summarise the main findings, which we contextualise in relation to the existing literature in the remainder of this section.

8.1 Temporal Structure Was Consistently Perceived in Tonal and Atonal Music

Our first research question was whether temporal structure is perceived in tonal and atonal music, and whether the effect of temporal degradation interacts with tonality. Consistently across non-musicians and musicians, elbow-function analysis revealed a higher plateau for tonal music than atonal music, with

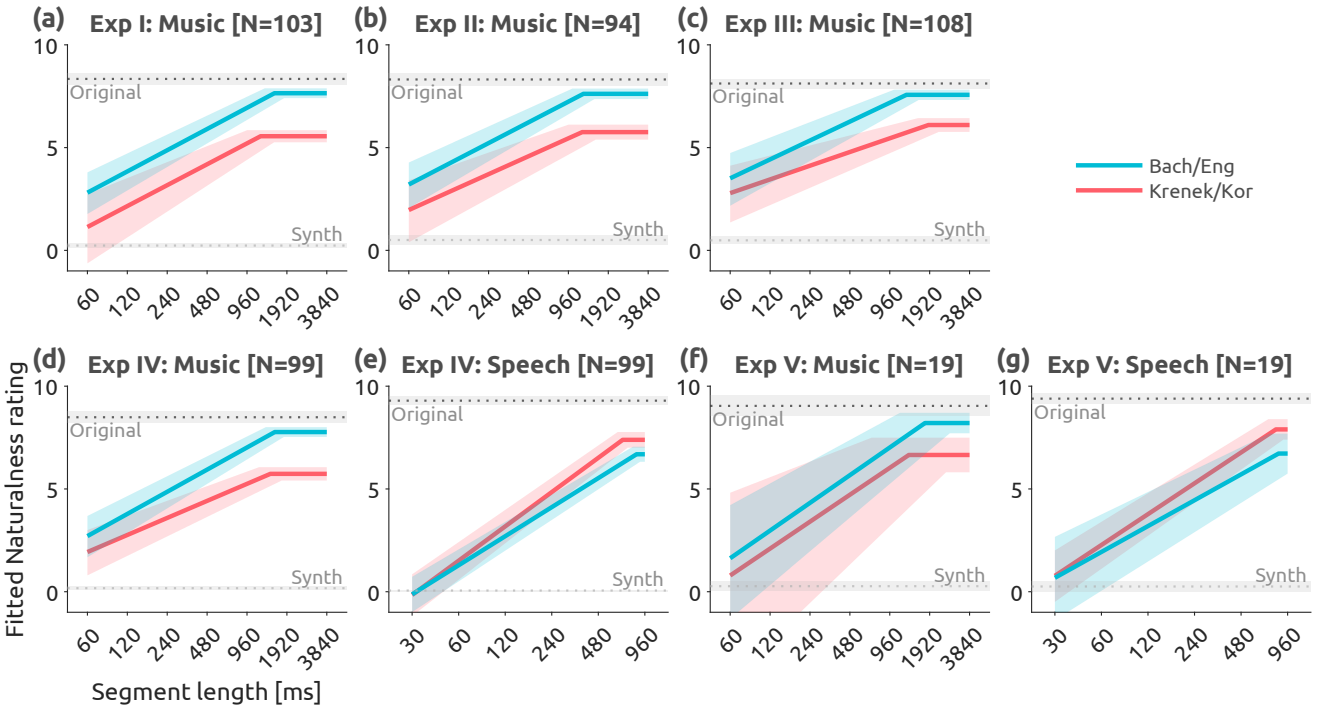

Figure 11: Fitted elbow functions from all experiments. Note that the naturalness responses to music samples in Experiment IV is a subset of a union of Experiment I and II.

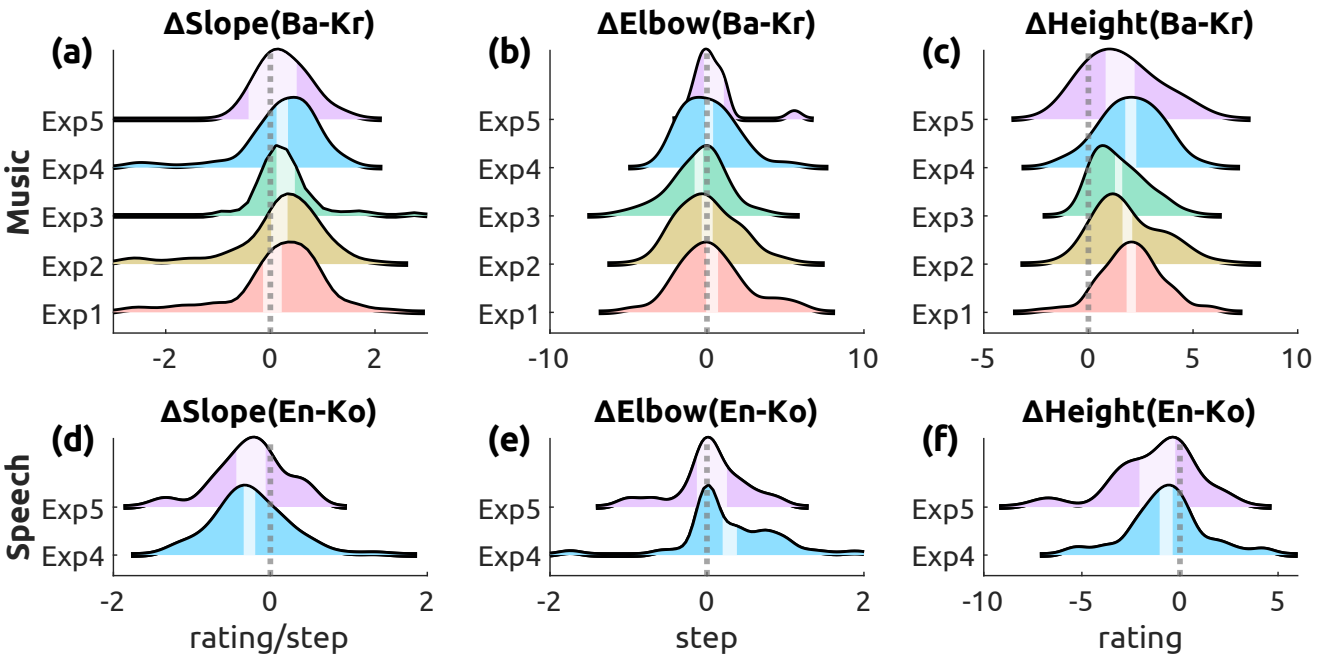

Figure 12: Ridgeline plots of differences in fitted parameters. Experiment I ($N = 103$), Experiment II ($N = 94$), Experiment III ($N = 108$), Experiment IV ($N = 99$), Experiment V ($N = 19$). Note that the naturalness responses to music samples in Experiment IV is a subset of a union of Experiment I and II.

similar slopes and elbow points. This indicates that the amount of temporal information contributed similarly to naturalness ratings for both tonal and atonal music, whilst the different plateaus reflect a main effect of tonality. To confirm this main effect, paired differences between tonal and atonal music were compared at each segment length; only a few points specific to certain exemplar sets were non-significant (Exp I: $P_{\mathrm{Bon}} = 0.249$ at 120 ms; Exp II: $P_{\mathrm{Bon}} = 0.216$ at 120 ms; Exp III: $P_{\mathrm{Bon}} = 1.0$ at 60 ms; all other comparisons: $P_{\mathrm{Bon}} < 10^{-14}$).

Emotional effects of music within a local context ($\sim$6 s), with global context disrupted, have been demonstrated in terms of perceived musical expressiveness and coherence at a level comparable to that of original excerpts (Tillmann & Bigand, 2001). Consistently with this, the current study found that even a shorter context (3.84 s) was rated similarly to the original versions (11.52 s), confirming that music can evoke emotional responses within brief temporal windows.

For atonal music, Lalitte et al. (2009) demonstrated that carefully re-composed atonal versions of Beethoven's piano sonatas (No. 21, Op. 53 and No. 17, Op. 31/2), with temporal structures preserved, can evoke a sense of arousal, suggesting an emotional role for formal functions (e.g., presentational, developmental, transitional, and closure sections in the sonata form) that is independent of tonal function. Krenek's Violin Sonata No. 2 is noted for adhering to the classical sonata form while employing a twelve-tone row (Pasler, 1985). Indeed, naturalness ratings increased with the length of the preserved local context in both tonal and atonal music.

The novelty of the current findings lies in the comparable contribution of temporal information to overall perception in tonal and atonal music within short ($\leq$3.84 s) contexts, alongside a consistently higher plateau for tonal music across experiments. However, given evidence for effects of longer contexts (up to $\sim$40 s; Farbood et al. 2015), it remains possible that tonal and atonal music diverge at timescales beyond those examined here. This should be examined empirically in future work.

8.2 Temporal Perception in Music Was Not Strongly Related to Speech

The second research question was whether temporal integration in music is a domain-specific or domain-general process. To address this question, we compared the effects of temporal degradation on naturalness ratings for familiar (English) versus unfamiliar (Korean) speech, in a manner analogous to the comparison of tonal (Bach) versus atonal (Krenek) music.

An unexpected finding was that the effect of familiarity is reversed in speech: unlike tonal music, which showed a higher plateau than atonal music, the native language showed a lower plateau than the foreign

language. We speculated that this could reflect higher sensitivity to disruptions in the native language than in tonal music among non-musicians. However, musicians showed a similarly reversed effect of familiarity in speech.

Interestingly, this familiarity effect was not strongly correlated between music and speech across individuals—in either non-musicians or musicians. This may reflect domain-specificity in music and/or speech processing. The hypothesis of a shared mechanism underlying high-level temporal structure perception in both domains could be tested further using online psychoacoustic measures (e.g., gap detection threshold; MacIntyre and Scott 2022). It also remains possible that ecologically salient sounds such as voice are processed differently from other complex sounds. Taken together, the current findings suggest that naturalness ratings may have reflected different perceptual dimensions depending on the type of information available: linguistic knowledge in speech, and tonal knowledge in music.

8.3 Musical Training Sensitised Temporal Perception in Atonal Music

The third research question was whether temporal perception in music is modulated by musical training. One major motivation for recruiting musicians trained in Western classical music was to test whether heightened familiarity with tonal music would modulate naturalness ratings, in relation to our speculation about the reversed familiarity effect in speech. Whilst we did not find any altered effect of tonality in music, we instead found that musicians showed higher slopes and plateau heights for atonal music than non-musicians. This suggests that musicians perceived atonal music as more natural when temporal structure was largely intact (higher plateau) and responded more sensitively to its degradation (higher slope).

We did not obtain information about musicians' familiarity with atonal music. However, it is common for musicians to acquire at least minimal exposure to atonal music through formal music education. Moreover, given that Krenek's composition closely follows the classical sonata form, trained musicians may have recognised and utilised the local temporal structure of the atonal piece more readily than non-musicians, resulting in higher naturalness ratings.

Beyond familiarity and musicological knowledge, prior studies have reported heightened sensory processing of temporal structures in musicians, including greater auditory temporal-interval discrimination (Banai et al., 2012), higher auditory temporal acuity for auditory fusion (Rammsayer & Altenmüller, 2006; Vibell et al., 2021), higher sensitivity to temporal fine structure (Bianchi et al., 2019; Mishra et al., 2015), and better performance on other psychoacoustic tasks (Carey et al., 2015). However, such heightened processing did not substantially account for individual differences in temporal structure perception of

speech or music in the current study. In principle, a higher temporal resolution in musicians would predict steeper rating degradation as temporal context is disrupted. Whilst Figure 9 shows numerically higher slopes for Bach in musicians, the effect was not significant. We consider this inconclusive given the small musician sample size.

8.4 Self-reported Perceptual Skill Associated with Temporal Perception

The last research question was whether individual differences in temporal perception are associated with self-reported musicality in nonmusicians. The most intriguing finding of the current study is that the self-reported Perceptual Abilities subscale was positively associated with the elbow point for Bach quilts and the difference in plateau heights between Bach and Krenek. These associations were replicated via cross-validation. The Perceptual Abilities subscale includes items such as "I am able to judge whether someone is a good singer or not" and "I usually know when I'm hearing a song for the first time."

The validity of self-report measures is generally considered to vary considerably depending on the nature of the questionnaire (Paulhus, Vazire, et al., 2007). However, for perceptual abilities specifically, self-reported measures have shown good agreement with behavioural measures in vision (Whillans & Nazroo, 2014), hearing (Kamil et al., 2015), and taste (Soter et al., 2008). In particular, the behavioural validity of the self-reported Perceptual Abilities subscale from Gold-MSI has been established in the original paper (Müllensiefen et al., 2014) and independently replicated in an online behavioural study (Correia et al., 2022). More broadly, a recent genome-wide association study leveraged surprisingly accurate self-reports of beat synchronisation ability (Niarchou et al., 2022), which correlated strongly with behavioural performance in a subset of participants. It is therefore anticipated that the current finding will replicate with behavioural measures in future work.

Which particular components of the Perceptual Abilities subscale are associated with heightened sensitivity to temporal structures in tonal music remains unclear. As an exploratory analysis (Supplementary Result S2.3), item-wise correlations were calculated. Items such as "Spotting Mistakes in Performance", "Recognising Familiar Tune", and "Judge Own Tonal Perception" showed stronger correlations than others. These may reflect more attentive listening, which is of particular interest given that a higher elbow point would also indicate a longer effective attention span in processing musical sounds—a capacity necessary for recognising musical phrases (Knösche et al., 2005).

8.5 Temporal Processing and Musical Emotion Mediated by Self-reported Musicality

Although the current study does not focus exclusively on musical emotion, it has several implications for understanding how music evokes affective responses. As noted above, prior research has shown that even a local context of ∼6 s can evoke percepts of musical expressiveness and coherence (Tillmann & Bigand, 2001). Additionally, a magnetoencephalography study demonstrated high sensitivity of the human auditory cortex to the emergence of temporal coherence (Kim et al., 2020). The present study further underscores the importance of local temporal structures spanning tens of milliseconds to several seconds in the general perception of music, with potential implications for evoked emotions and musical pleasure.

Moreover, the study reveals substantial intersubject variability in perceiving temporal structures, mirroring findings in musical reward (Mas-Herrero et al., 2012), which has been linked to diffusion-based properties of ventral white matter tracts connecting the auditory cortex and cortical limbic areas (Martínez-Molina et al., 2016). The observed association between musical sophistication and sensitivity to temporal structures suggests a possible involvement of musical reward. Reward may sharpen perception: in vision, evidence suggests that high reward reduces internal noise and enhances perceptual learning (Tamaki et al., 2020); in hearing, higher reward has been shown to reduce speech perception thresholds (Bianco et al., 2021); and in a rat model, reward heightened frequency-specific gain in the primary auditory cortex (Hui et al., 2009). Indeed, a recent study found that musical anhedonia was associated with generally lower Gold-MSI subscale scores, including Perceptual Abilities (Kathios et al., 2024). Thus, individuals sensitive to musical reward may be more likely to develop greater perceptual acuity for musical structures, including temporal structure, even without formal training. It is important to note, however, that this link remains tentative in the current study, as we observed no association with Active Engagement—a Gold-MSI subscale encompassing items related to time and financial investment in musical activities, and thus more closely tied to musical reward.

Finally, predictive coding accounts of musical emotion typically assume the recognition of discrete events, and experiments often present chords or notes in clear separation. However, real-world music frequently contains subtle onsets, prosodic variation in timbre, and dynamic local tempi that enhance emotional responses (Chapin et al., 2010) whilst posing challenges for the discretisation of musical events. Segmentation and parsing in music perception therefore warrant further investigation to better understand how real-world music evokes pleasure (Sridharan et al., 2007; Williams et al., 2022).

8.6 Limitations

Methodological limitations of the current study should be acknowledged. First, there are well-documented concerns about the quality of data collected from online platforms such as MTurk (Douglas et al., 2023; Thomas & Clifford, 2017). To address this, we introduced a semi-automatic post-screening procedure based on objective criteria, systematically excluding $\sim$13% of total responses, and replicated the general findings in a subset of participants recruited from local in-person communities. Despite these concerns, we believe the data offer valuable empirical insights into music perception.

Second, the coverage of musical samples was narrow. Rather than sampling broadly across many excerpts (e.g., Cowen et al., 2020), we manually selected high-quality exemplars on musicological grounds and manipulated stimuli to precisely control temporal information, which limited us to two compositions from Western classical music. Future research could explore the interaction (or parallelism) of tonal and temporal structures using large-scale musical corpora spanning diverse genres, with algorithmic quantification of tonality combined with the quilting algorithm.

Third, the behavioural rating of "naturalness" may have reflected multiple perceptual dimensions, including sensory judgements about the sound, cognitive judgements about structural continuity, and affective responses. Reference examples (Bach original and Bach quilts) were provided—without any explicit instruction on which should be considered "natural"—prior to the main blocks to anchor participants' use of the scale, potentially based on the quilting manipulation. However, the subcomponents of the naturalness judgement remain to be characterised. It also remains possible that the current design might have biased participants to rate any deviation from the reference stimulus as "unnatural", including stylistic variations. A future study with a broader range of musical styles would help address this limitation.

8.7 Conclusions

In conclusion, the current study demonstrates that tonal and temporal structures in music are processed in parallel. High sensitivity to temporal structures at short timescales was consistently found across both musicians and non-musicians and across multiple samples. Musicality—in both musicians and non-musicians—was found to be associated with temporal structure perception. Comparison with native and foreign speech perception revealed a complex and domain-specific nature of temporal processing. Taken together, the current study underscores the significance of local temporal structures in music perception.

Data and code availability statement

Exemplar audio stimuli, de-identified (participant IDs were hashed with a secured secret key) data, and relevant MATLAB code are available at Open Science Framework (https://osf.io/e58tr/; view-only link for review: https://osf.io/e58tr/overview?view_only=3babaeff6f764f8184e6b82a060cac1a). An executable virtual environment with all data and code is publicly available at Code Ocean (https://codeocean.com/capsule/5141283/).

Acknowledgements

This work was in part supported by Charles Lafitte Foundation Program in Psychological Research at Duke University to SGK, Dean's Summer Research Fellowship at Duke University to YY, and National Institutes of Health grant R21DC016386 to TO. We thank Dr. Scott Lindroth for discussion and suggestions and to Dr. Jiayue Liu for technical consultations.

CRediT author contribution statement

SGK: Conceptualisation, Funding acquisition, Investigation, Data curation, Methodology, Formal Analysis, Software, Visualisation, Validation, Writing–original draft, Writing–review & editing; YY: Funding acquisition, Investigation, Project administration, Writing–review & editing; DM: Methodology, Formal Analysis, Writing–review & editing; TO: Conceptualisation, Funding acquisition, Methodology, Resources, Supervision, Writing–review & editing.

References

Banai, K., Fisher, S., & Ganot, R. (2012). The effects of context and musical training on auditory temporal-interval discrimination. *Hearing research*, *284*(1-2), 59–66.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Bianchi, F., Carney, L. H., Dau, T., & Santurette, S. (2019). Effects of musical training and hearing loss on fundamental frequency discrimination and temporal fine structure processing: Psychophysics and modeling. *Journal of the Association for Research in Otolaryngology*, *20*(3), 263–277.

Bianco, R., Mills, G., de Kerangal, M., Rosen, S., & Chait, M. (2021). Reward enhances online participants' engagement with a demanding auditory task. *Trends in Hearing*, *25*, 23312165211025941.

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021). Music-selective neural populations arise without musical training. *Journal of Neurophysiology*.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, *137*, 81–105.

Chapin, H., Jantzen, K., Kelso, J. S., Steinberg, F., & Large, E. (2010). Dynamic emotional and neural responses to music depend on performance expression and listener experience. *PloS one*, *5*(12), e13812.

Cheung, V. K., Harrison, P. M., Meyer, L., Pearce, M. T., Haynes, J.-D., & Koelsch, S. (2019). Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology*, *29*(23), 4084–4092.

Correia, A. I., Vincenzi, M., Vanzella, P., Pinheiro, A. P., Lima, C. F., & Schellenberg, E. G. (2022). Can musical ability be tested online? *Behavior Research Methods*, *54*(2), 955–969.

Cowen, A. S., Fang, X., Sauter, D., & Keltner, D. (2020). What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences*, *117*(4), 1924–1934.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration. *The Journal of the Acoustical Society of America*, *102*(5), 2906–2919.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *PLOS ONE*, *18*(3), 1–17.

Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., & Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in neuroscience*, *9*, 157.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, *364*(1521), 1211–1221.

Gingras, B., Honing, H., Peretz, I., Trainor, L. J., & Fisher, S. E. (2015). Defining the biological bases of individual differences in musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664).

691 Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A., & Zatorre, R. J. (2019). Predictability and
692 uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience*, *39*(47),
693 9397–9409.
694 Harrison, P. M., Bianco, R., Chait, M., & Pearce, M. T. (2020). Ppm-decay: A computational model of
695 auditory prediction with memory decay. *PLoS computational biology*, *16*(11), e1008304.
696 Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning:*
697 *Data mining, inference, and prediction* (Vol. 2). Springer.
698 Hui, G. K., Wong, K. L., Chavez, C. M., Leon, M. I., Robin, K. M., & Weinberger, N. M. (2009).
699 Conditioned tone control of brain reward behavior produces highly specific representational gain
700 in the primary auditory cortex. *Neurobiology of learning and memory*, *92*(1), 27–34.
701 Huron, D. B. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT press.
702 Kamil, R. J., Genther, D. J., & Lin, F. R. (2015). Factors associated with the accuracy of subjective
703 assessments of hearing impairment. *Ear and hearing*, *36*(1), 164–167.
704 Kathios, N., Patel, A. D., & Loui, P. (2024). Musical anhedonia, timbre, and the rewards of music
705 listening. *Cognition*, *243*, 105672.
706 Kim, S.-G., De Martino, F., & Overath, T. (2024). Linguistic modulation of the neural encoding of
707 phonemes. *Cerebral Cortex*, *34*(4), bhae155.
708 Kim, S.-G., Mueller, K., Lepsien, J., Mildner, T., & Fritz, T. H. (2019). Brain networks underlying
709 aesthetic appreciation as modulated by interaction of the spectral and temporal organisations of
710 music. *Scientific Reports*, *9*(1), 19446.
711 Kim, S.-G., Poeppel, D., & Overath, T. (2020). Modulation change detection in human auditory cortex:
712 Evidence for asymmetric, non-linear edge detection. *European Journal of Neuroscience*, *52*(2),
713 2889–2904.
714 Knösche, T. R., Neuhaus, C., Haueisen, J., Alter, K., Maess, B., Witte, O. W., & Friederici, A. D. (2005).
715 Perception of phrase structure in music. *Human Brain Mapping*, *24*(4), 259–273.
716 Lad, M., Billig, A. J., Kumar, S., & Griffiths, T. D. (2022). A specific relationship between musical
717 sophistication and auditory working memory. *Scientific reports*, *12*(1), 3517.
718 Lalitte, P., Bigand, E., Kantor-Martynuska, J., & Delbé, C. (2009). On listening to atonal variants of
719 two piano sonatas by beethoven. *Music Perception*, *26*(3), 223–234.
720 Levitin, D. J., & Menon, V. (2003). Musical structure is processed in "language" areas of the brain: A
721 possible role for brodmann area 47 in temporal coherence. *Neuroimage*, *20*(4), 2142–2152.
722 MacIntyre, A. D., & Scott, S. K. (2022). Listeners are sensitive to the speech breathing time series:
723 Evidence from a gap detection task. *Cognition*, *225*, 105171.

Martínez-Molina, N., Mas-Herrero, E., Rodríguez-Fornells, A., Zatorre, R. J., & Marco-Pallarés, J. (2016). Neural correlates of specific musical anhedonia. *Proceedings of the National Academy of Sciences*, *113*(46), E7337–E7345.

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2012). Individual differences in music reward experiences. *Music Perception: An Interdisciplinary Journal*, *31*(2), 118–138.

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., et al. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868.

Meyer, L. B. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, *15*(4), 412–424.

Mishra, S. K., Panda, M. R., & Raj, S. (2015). Influence of musical training on sensitivity to temporal fine structure. *International journal of audiology*, *54*(4), 220–226.

Morel, P. (2018). Gramm: Grammar of graphics plotting in matlab. *Journal of Open Source Software*, *3*(23), 568.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS one*, *9*(2), e89642.

Niarchou, M., Gustavson, D. E., Sathirapongsasuti, J. F., Anglada-Tort, M., Eising, E., Bell, E., McArthur, E., Straub, P., McAuley, J. D., et al. (2022). Genome-wide association study of musical beat synchronization demonstrates high polygenicity. *Nature Human Behaviour*, *6*(9), 1292–1309.

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology*, *32*(7), 1470–1484.

Norman-Haignere, S. V., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *neuron*, *88*(6), 1281–1296.

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., et al. (2022a). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature human behaviour*, *6*(3), 455–469.

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., et al. (2022b). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature human behaviour*, *6*(3), 455–469.

Norman-Haignere, S. V., & McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS biology*, *16*(12), e2005127.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature neuroscience*, *18*(6), 903–911.

Overath, T., & Paik, J. H. (2021). From acoustic to linguistic analysis of temporal speech structure: Acousto-linguistic transformation during speech perception using speech quilts. *Neuroimage*, *235*, 117887.

Overath, T., Zhang, Y., Sanes, D. H., & Poeppel, D. (2012). Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: Fmri evidence. *Journal of neurophysiology*, *107*(8), 2042–2056.

Pasler, J. (1985). Ernst křenek-in retrospect. *Perspectives of New Music*, 424–432.

Paulhus, D. L., Vazire, S., et al. (2007). The self-report method. *Handbook of research methods in personality psychology*, *1*(2007), 224–239.

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the new York Academy of Sciences*, *1423*(1), 378–395.

Peretz, I., & Hyde, K. L. (2003). What is specific to music processing? insights from congenital amusia. *Trends in cognitive sciences*, *7*(8), 362–367.

Rammsayer, T., & Altenmüller, E. (2006). Temporal information processing in musicians and nonmusicians. *Music Perception*, *24*(1), 37–48.

Rimmele, J. M., Kern, P., Lubinus, C., Frieler, K., Poeppel, D., & Assaneo, M. F. (2022). Musical sophistication and speech auditory-motor coupling: Easy tests for quick answers. *Frontiers in neuroscience*, *15*, 764342.

Soter, A., Kim, J., Jackman, A., Tourbier, I., Kaul, A., & Doty, R. L. (2008). Accuracy of self-report in detecting taste dysfunction. *The Laryngoscope*, *118*(4), 611–617.

Sridharan, D., Levitin, D. J., Chafe, C. H., Berger, J., & Menon, V. (2007). Neural dynamics of event segmentation in music: Converging evidence for dissociable ventral and dorsal networks. *Neuron*, *55*(3), 521–532.

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1517.

Tamaki, M., Berard, A. V., Barnes-Diana, T., Siegel, J., Watanabe, T., & Sasaki, Y. (2020). Reward does not facilitate visual perceptual learning until sleep occurs. *Proceedings of the National Academy of Sciences*, *117*(2), 959–968.

792 Thomas, K. A., & Clifford, S. (2017). Validity and mechanical turk: An assessment of exclusion methods
793 and interactive experiments. *Computers in Human Behavior*, *77*, 184–197.
794 Tillmann, B., & Bigand, E. (1996). Does formal musical structure affect perception of musical expres-
795 siveness? *Psychology of music*, *24*(1), 3–17.
796 Tillmann, B., & Bigand, E. (2001). Global context effect in normal and scrambled musical sequences.
797 *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1185.
798 Vibell, J., Lim, A., & Sinnett, S. (2021). Temporal perception and attention in trained musicians. *Music*
799 *Perception: An Interdisciplinary Journal*, *38*(3), 293–312.
800 Vuust, P., Heggli, O. A., Friston, K. J., & Kringelbach, M. L. (2022). Music in the brain. *Nature Reviews*
801 *Neuroscience*, *23*(5), 287–305.
802 Whillans, J., & Nazroo, J. (2014). Assessment of visual impairment: The relationship between self-
803 reported vision and 'gold-standard'measured visual acuity. *British Journal of Visual Impairment*,
804 *32*(3), 236–248.
805 Williams, J. A., Margulis, E. H., Nastase, S. A., Chen, J., Hasson, U., Norman, K. A., & Baldassano,
806 C. (2022). High-order areas and auditory cortex both represent the high-level event structure of
807 music. *Journal of Cognitive Neuroscience*, *34*(4), 699–714.
808 Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate
809 web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072.

Supplementary Information for "Temporal Structure Perception in Tonal and Atonal Music: An Online Behavioural Study"

Seung-Goo Kim[1,2,a,*], Daniel Müllensiefen[3,b], Ying Yu[1,c], Tobias Overath[1,d]

[1]Duke University, Durham, North Carolina, U.S.A.

[2]Max Planck Institute for Empirical Aesthetics, Germany

[3]University of Hamburg, Hamburg, Germany

[a]0000-0003-0558-8547

[*]Correspondence: seung-goo.kim@ae.mpg.de

April 10, 2026

S1 Supplementary Methods

S1.1 Musical piece selection

Our selection criteria and reasoning were as follows: (a) solo pieces: when used to create quilts, polyphonic/homophonic music (ensembles and even piano solos) resulted in obvious artefacts arising from multiple parts. Thus, we decided solo pieces would be more feasible to manipulate without introducing excessive artefacts. (b) 12-tone serialism otherwise traditionally composed: because we wanted to use atonal music as a control for the tonal music while varying only the tonal structure, we chose a piece composed in 12-tone serialism (and thus presenting no distinctive tonal centre) but without other distinctive features (e.g., non-traditional bowing, sparse notes) and with a similar level of virtuosity. Krenek's Sonata for Solo Violin No. 2 is considered one of the strictest 12-tone serialism compositions (Pasler, 1985). (c) the same artist: to minimise acoustic differences between tonal and atonal music, we sought recordings performed by the same artist (preferably using the same instrument and recorded in a similar

environment with similar mastering). There are many other great atonal pieces, but the Bach–Krenek pair was one of very few pieces that met all three criteria. Selected music data were recordings of Johann Sebastian Bach's Violin Sonatas and Partitas (2018, Audite Musikproduktion, Detmold, Germany[1]) and Ernst Krenek's Sonata for Solo Violin No. 2 (2013, Audite Musikproduktion, Detmold, Germany[2]), both performed by Christoph Schickedanz. Using an R wrapper for Spotify's Web API[3], tempi of the tracks were retrieved from the Spotify's database[4]. For Bach (32 movements), mean BPM $= 108.70 \pm 29.62$. For Krenek (3 movements), mean BPM $= 98.68 \pm 5.49$.

S1.2 Temporal structure manipulation using the quilting algorithm

The quilting algorithm reorders segments of a given length to effectively disrupt temporal structures longer than the segment length while minimising artefacts in short-term acoustics introduced by the process. This is achieved by selecting the order of segments that minimises acoustic changes and by concatenating segments after aligning phases. The following describes the details of each step.

First, the stereo recording was averaged across the left and right channels and down-sampled to 20 kHz with an anti-aliasing high-pass filter. The quilting algorithm tends to concatenate all silent segments. To avoid this artefact, silent gaps (< -40 dB relative to the maximum RMS for 60 ms [i.e., two 30-ms segments]) were removed from each movement (on average 8.16 ± 4.87 s $[3.47\% \pm 2.72\%]$ per Bach movement and 30.77 ± 8.37 s $[17.28\% \pm 1.07\%]$ per Krenek movement).

Subsequently, the whole music signal (32 Bach movements [totalling 2 hours 8.9 min] and three Krenek movements [totalling 8.8 min]) was further divided into 50-s excerpts to constrain scrambling. For each 50-s excerpt, the audio samples were segmented at a given length—either 60, 120, 240, 480, 960, 1820 or 3640 ms (i.e., log-linearly spaced). From the perspective of a cochlear model with a short-term (60-ms) integration window, the changes introduced by reordering segments can be quantified by calculating an L_2-norm distance between the 30-ms boundaries of segments in the cochleogram as:

$$\Delta_{i \to j} = \left\| \sum_{t \in \mathcal{A}_i} \mathbf{c}(t) - \sum_{t \in \mathcal{B}_j} \mathbf{c}(t) \right\|_2^2, \tag{S1}$$

[1]https://music.apple.com/us/album/bach-sei-solo-%C3%A1-violino-senza-basso-accompagnato/1394469696

[2]https://music.apple.com/us/album/ernst-krenek-works-for-violin-sonata-for-solo-violin/594802051

[3]https://github.com/charlie86/spotifyr

[4]The feature retrieval via Spotify's API is no longer available for new users since 27 November 2024. See: https://developer.spotify.com/blog/2024-11-27-changes-to-the-web-api

where $\mathbf{c}(t)$ is the vector of the cochleogram spectrum at time t, $\mathcal{A}_i$ is the set of time points corresponding to the last 30 ms of the i-th segment, and $\mathcal{B}_j$ is the set of time points corresponding to the first 30 ms of the j-th segment. Once an initial segment is chosen (say the i-th segment), the quilting algorithm selects the next segment (e.g., the j-th segment). Importantly, it does not attempt to minimise $\Delta_{i\to j}$ itself but instead aims to preserve the original transition ($\Delta_{i\to i+1}$) by minimising δ:

$$\arg\min_{j\in\mathcal{S}_{-i}} \delta(i,j) \;\equiv\; \arg\min_{j\in\mathcal{S}_{-i}} \|\Delta_{i\to i+1} - \Delta_{i\to j}\|_2^2\,, \tag{S2}$$

where the set of candidate segments $\mathcal{S}_{-i}$ excludes the segment itself (the i-th) and the original consecutive segment (the $i{+}1$-th).

Now, when the algorithm finds N transitions for a sequence with $N+1$ segments, the global cost function can be defined by the mean absolute difference (MAD):

$$\mathrm{MAD} = \frac{1}{N}\sum_{i,j\in\mathcal{K}} \delta(i,j) \tag{S3}$$

for a set $\mathcal{K}$ of selected segments. Because of the exhaustive search (i.e., the algorithm computes absolute differences for all possible candidates), the algorithm is deterministic except for the initialisation. That is, the entire sequence of segments and its cost (i.e., MAD) are functions of the initialisation, which was drawn from a uniform distribution in the original algorithm (Overath et al., 2015). In the present study, we also exhaustively searched all possible initialisations (except for the last segment, as the following segment is undefined) and found a global minimum of the cost function (i.e., the minimal MAD across all possible initialisations). On a Linux workstation with 16 cores of Intel Xeon 3.6 GHz and 64 GB RAM, the entire process took 2 days and 9 hours. The MAD is zero in the original recordings. An average MAD across all possible initialisations was $1.08\times10^{-3}\pm7.65\times10^{-4}$ for Bach and $1.47\times10^{-3}\pm9.84\times10^{-4}$ for Krenek. The global minimum of MAD was $6.94\times10^{-4}\pm3.87\times10^{-4}$ (64% of the mean) for Bach and $8.70\times10^{-4}\pm4.80\times10^{-4}$ (59% of the mean) for Krenek, suggesting that the exhaustive search achieved much better quilting quality than would be expected by chance. After determining the optimally scrambled order of segments, 11.52-s quilts were concatenated using a phase-shift PSOLA algorithm and up-sampled to 44.1 kHz.

S1.3 Model-based stimuli matching

Although the Bach and Krenek recordings were performed by the same artist to minimise acoustic discrepancies between the two musical styles, considerable acoustic differences between the Bach and Krenek pieces remain because acoustic and musical characteristics are highly multicollinear in natural music (Broze & Huron, 2013). However, leveraging the large number of created quilts (952 Bach quilts and 56 Krenek quilts), we sought to find pairs of music quilts that were as acoustically similar as possible in terms of peripheral and central auditory neural responses (Norman-Haignere & McDermott, 2018). Specifically, we calculated four statistical moments of cochleograms (i.e., a biologically constrained spectrogram that mimics cochlear neuronal activity) and of cortical representations (i.e., biologically constrained spectrotemporal modulation filters that mimic primary auditory neuronal activity) along the temporal dimension, collapsing the temporal dimension while preserving spectral bins for the cochleogram and spectral bins and modulation rate bins for the cortical representation.

We calculated Pearson correlations between the vectorised moments of Bach and Krenek quilts and then averaged all correlation coefficients for each pair of Bach and Krenek quilts. For 1,088 pairs, the mean correlation was 0.53 ± 0.06, with a range of $[0.27, 0.76]$. We selected the top eight pairs for each segment length to use as 'music quilts' in the experiments (mean $r = 0.64 \pm 0.04$). The selected quilts were drawn from 16 Bach movements and all three Krenek movements. Further details of the selected quilts are provided in the Supplementary File.

S1.4 Perceived loudness equalisation

For all created stimuli, the perceived loudness based on ISO standard 523-1 (Zwicker's method) in sones (i.e., perceived loudness) was converted to phons to compute a gain factor, as implemented in the MATLAB Audio Toolbox (R2021a v9.10.0.1602886; `acousticLoudness.m` and `sone2phon.m`), as:

$$\text{Gain} = 10^{(90-\text{Phon})/20} \tag{S4}$$

The equalised loudness of the music quilts was on average 31.07 ± 0.24 sones (minimum $= 30.43$, maximum $= 31.60$; the original recordings were about 40 sones). While the actual loudness depended on the individual volume settings of the online participants' devices, a brief calibration during the headphone screening was intended to adjust it to around 60 dB SPL.

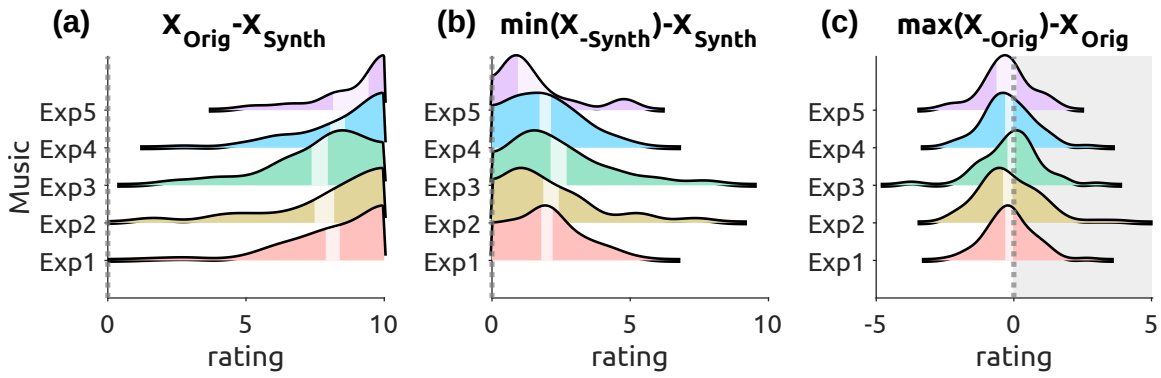

Figure S1: Ridgeline plots of rating metrics. (a) Differences between the Original and Synth stimuli. (b) Differences between the minimal rating and the Synth. (c) Differences between the maximal rating and the Original. Light vertical bands on ridges demarcate 95% confidence intervals of the mean. Grey shading demarcates where the assumptions for normalisation do not hold. Grey vertical dashed lines indicate absolute null levels.

S1.5 Normalisation of ratings using reference stimuli

We decided not to normalise the raw rating value X ($X \in \mathbb{R},\ 0 \leq X \leq 10$) because it would have greatly changed the boundaries of the normalised value Y. Here we provide the details.

The reference stimuli ("Original" and "Synth") may be used to estimate the floor and ceiling levels of the naturalness rating for a given participant. That is, the average ratings of these reference stimuli may be used to normalise ratings:

$$Y_i = \frac{X_i - X_s}{X_o - X_s} \tag{S5}$$

where Y_i is the normalised rating for the i-th stimulus, ideally confined between 0 and 1; X_i is the raw rating for the i-th stimulus; and X_o and X_s are the raw ratings for the Original and Synth stimuli, respectively.

For the normalisation to allow meaningful comparisons across participants, the normalised value should satisfy: $Y \in \mathbb{R},\ 0 \leq Y \leq 1$. For this, the following constraints are required: (1) $X_o > X_s$ for $Y_i \in \mathbb{R}$, (2) $X_i \geq X_s,\ \forall i$ for $Y_i > 0$, and (3) $\max X_i \leq X_o,\ \forall i$ for $Y_i \leq 1$.

The first condition is fulfilled, as $X_o > X_s$ in all cases. The second condition is also fulfilled because we post-screened responses (Figure S1b). However, the third condition is not met for the majority of participants. That is, $\max X_i > X_o$ for some i: 63.81% in Experiment I; 59.00% in Experiment II; 53.39% in Experiment III; 63.30% in Experiment IV; and 78.26% in Experiment V (samples in the shaded regions in Figure S1c). Thus, we decided not to normalise the rating values.

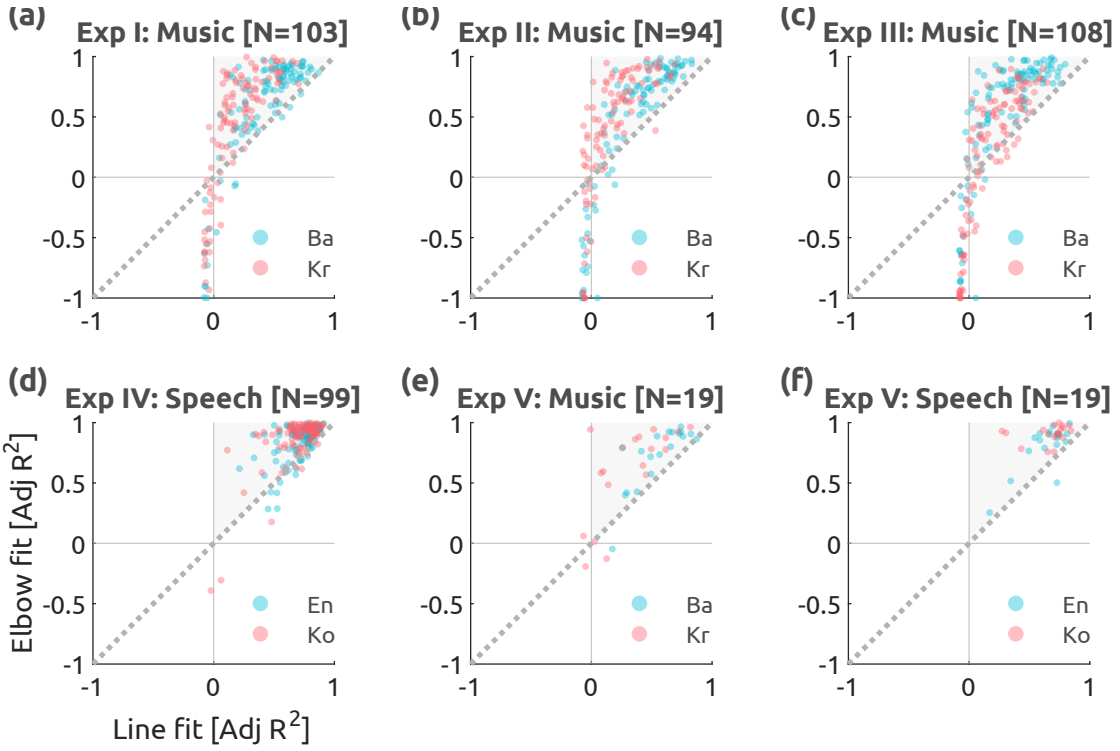

Figure S2: Fitting of elbow functions over linear functions. Abbreviations: Ba, Bach; Kr, Krenek; En, English; Ko, Korean.

S1.6 Elbow function vs. linear function

The elbow models were selected over linear models based on the adjusted R^2. The explained variance of the linear and elbow models is compared in Figure S2. Most responses (79% on average) were located above the identity line and to the right of the horizontal zero line (i.e., the upper triangular area in the first quadrant; shaded in grey), indicating improved fitting with an elbow function compared with a linear function. Individuals with negative adjusted R^2 for linear functions (i.e., data not well explained by linear functions) also showed even poorer fits with elbow functions.

In addition, motivated by the initial plateau observed in the first exemplar set of Bach quilts (on average, no increase from 60 ms to 120 ms), a logistic function,

$$f(x) = \frac{a}{1 + \exp\left(-b(x - c)\right)},$$

was also explored. However, as shown in Figure S3, the logistic fit was generally not better than that of the elbow function, except for the speech quilts (Exp. IV–Speech, Exp. V–Speech). Thus, we used the elbow function for the main analysis.

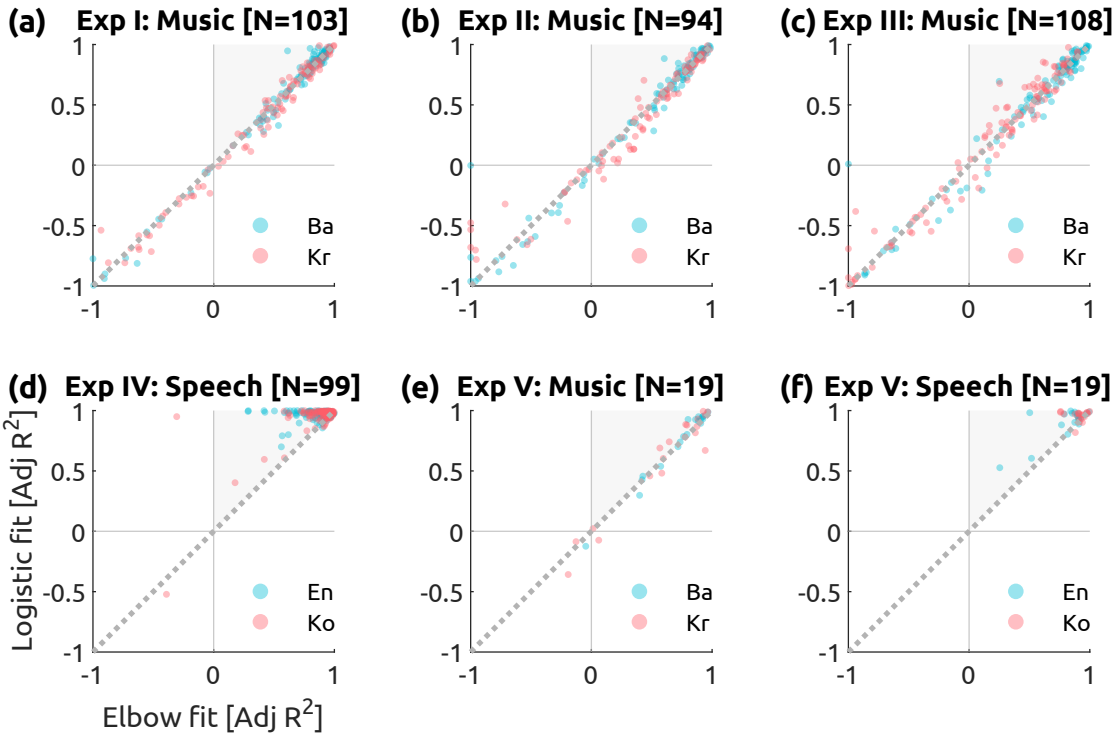

Figure S3: Fitting of logistic functions over elbow functions.

S2 Supplementary Results

S2.1 Sample demographics

Table S1 summarises the overall participants of all experiments. Table S2 lays out the distribution of age, gender and racial groups of the participants included in the reported analysis. When considering MTurk participants without overlap (i.e., Experiments I–IV; 423 responses), no significant differences in distributions were found (χ^2-test, $\min P = 0.137$). Compared with the MTurk participants, musicians recruited from university orchestras (Experiment V) were younger ($\chi^2 = 45.30$, $P < 10^{-3}$) and differed

Table S1: Overview of participants of all experiments. *Participants overlapped with Experiment I and III.

Experiment	Stimulus sets	Participant pool	Participated [N]	Analyzed [N]
I. Discovery	Music set A	Amazon MTurk	109	103
II. Replication	Music set A	Amazon MTurk	108	94
III. Generalization	Music set B	Amazon MTurk	123	108
IV. Cross-domain*	Speech	Amazon MTurk	120	99
V. Musicianship	Music set A and Speech	Local musicians	27	19

Table S2: Demographic distribution. Age Group: A1: 18-29, A2: 30-44, A3: 45-59, A4: 60+, A5: others; Gender Group: G1: female, G2: male, G3: others; Race Group: R1: European, R2: Asian, R3: African, R4: Others

Exp No.	Total	A1	A2	A3	A4	A5	G1	G2	G3	R1	R2	R3	R4
I	103	31	44	24	3	1	44	58	1	74	15	9	5
II	94	24	57	8	5	0	46	46	2	72	7	8	7
III	108	28	54	17	9	0	41	67	0	75	12	18	3
IV	99	26	47	20	5	1	45	53	1	75	11	5	8
V	19	16	3	0	0	0	13	5	1	9	5	1	4

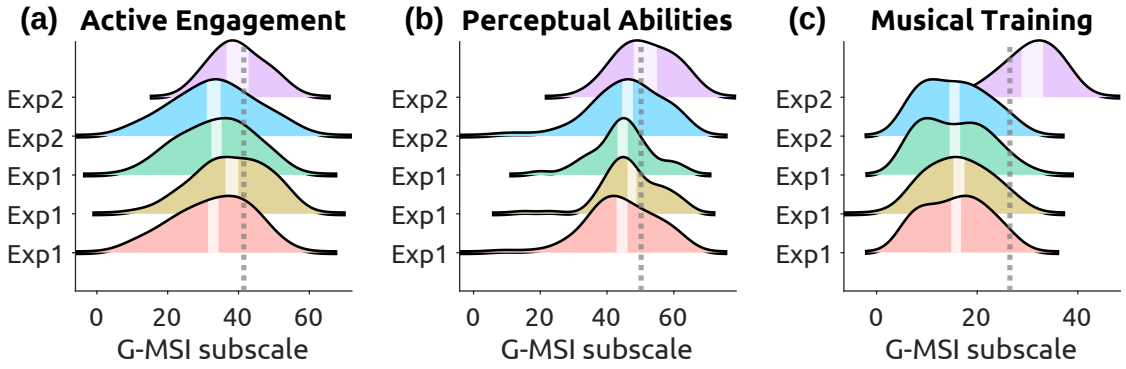

Figure S4: Ridgeline plots of Gold-MSI subscales. Each density plot represents Gold-MSI subscale scores [(a) Active Engagement, (b) Perceptual Abilities, (c) Musical Training] for each experiment. Light vertical bands on ridges demarcate 95% confidence intervals of the mean. Grey vertical dashed lines indicate normative means of subscales from a large-scale study (Müllensiefen et al., 2014). Note that the participants of Experiment IV are a subset of the union of Experiments I and II.

in racial composition ($\chi^2 = 25.64$, $P = 0.012$; due to the under-representation of individuals of African ancestry among musicians) but did not differ in terms of gender distribution ($P = 0.1043$).

Table S3 and Figure S4 display the means and 95% confidence intervals of the Goldsmith Musical Sophistication Index (Gold-MSI) subscales of the analysed participants. For reference, normative statistics from a large ($N = 147{,}633$) sample are also shown (Müllensiefen et al., 2014).

For participants who rated speech quilts, their language proficiencies and regions of origin were surveyed (Table S4). Among all analysed responses, all participants reported being native English speakers and having no knowledge of the Korean language. Most also reported no knowledge of other East Asian languages. The majority of participants originally came from the North American region.

Table S3: Musical Sophistication Metrics. Unequal sample size, unequal variance two-sample t-test performed with respect to the normative statistics (Müllensiefen et al., 2014). The normative means were subtracted from the observed means. Pooled degrees of freedom were calculated by the Welch–Satterthwaite equation. Uncorrected two-tailed P-values are given. Abbreviations: AE, Active Engagement; PA, Perceptual Abilities; MT, Musical Training; M2014, Müllensiefen et al., 2014.

Exp No.	N	GMSI	Mean	Stdev	CI	t	df	P
I	103	AE	32.786	10.485	[30.737, 34.836]	-8.45	102.14	$< 10^{-}12$
I	103	PA	44.612	10.042	[42.649, 46.574]	-5.65	102.09	$< 10^{-}6$
I	103	MT	15.660	6.059	[14.476, 16.844]	-18.17	102.51	$< 10^{-}33$
II	94	AE	37.979	9.880	[35.955, 40.002]	-3.47	93.13	$< 10^{-}3$
II	94	PA	47.457	8.265	[45.765, 49.150]	-3.22	93.11	0.002
II	94	MT	16.362	6.003	[15.132, 17.591]	-16.39	93.43	$< 10^{-}28$
III	108	AE	33.611	10.123	[31.680, 35.542]	-8.12	107.16	$< 10^{-}12$
III	108	PA	44.630	8.469	[43.014, 46.245]	-6.83	107.13	$< 10^{-}9$
III	108	MT	15.463	6.388	[14.244, 16.682]	-17.97	107.50	$< 10^{-}33$
IV	99	AE	33.222	11.233	[30.982, 35.463]	-7.35	98.11	$< 10^{-}10$
IV	99	PA	46.071	10.428	[43.991, 48.151]	-3.94	98.07	$< 10^{-}3$
IV	99	MT	15.465	6.130	[14.242, 16.687]	-17.92	98.46	$< 10^{-}32$
V	19	AE	39.789	6.941	[36.444, 43.135]	-1.09	18.01	0.292
V	19	PA	51.526	7.799	[47.767, 55.285]	0.74	18.00	0.468
V	19	MT	30.895	4.920	[28.523, 33.266]	3.87	18.03	0.001
M2014	147633	AE	41.520	10.360	[41.467, 41.573]	N/A	N/A	N/A
M2014	147633	PA	50.200	7.860	[50.160, 50.240]	N/A	N/A	N/A
M2014	147633	MT	26.520	11.440	[26.462, 26.578]	N/A	N/A	N/A

Table S4: Language background distribution. Language Proficiency Group: M1=None in Mandarin, M2=Elementary in Mandarin, M3=Native Mandarin; C1=None in Cantonese, C2=Elementary in Cantonese; J1=None in Japanese, J2=Elementary in Japanese, J3=Native Japanese; Region of Origin Group: O1: North America, O2: Central or South America, O3: European, O4: Asian.

Exp No.	Total	M1	M2	M3	C1	C2	J1	J2	J3	O1	O2	O3	O4
IV	99	99	94	4	1	97	2	92	6	92	3	2	2
V	19	19	9	5	5	17	2	17	2	17	0	0	2

Table S5: Summary of analysis of variance (ANOVA) models. FDR-corrected $P < 0.05$ are marked in bold. Abbreviations: df: degrees of freedom, η_p^2: partial eta-squared, Adj. R^2: adjusted R-squared.

Factor	Metric	F	df	η_p^2	P	P_{FDR}	Adj. R^2
age	Slope (Bach)	3.649	[4,304]	0.047	0.006	0.051	0.040
age	Elbow (Bach)	1.801	[4,304]	0.024	0.129	0.309	-0.003
age	Height (Bach)	0.469	[4,304]	0.006	0.758	0.867	-0.026
age	Slope (Krenek)	0.568	[4,304]	0.008	0.686	0.867	-0.022
age	Elbow (Krenek)	1.413	[4,304]	0.019	0.230	0.501	0.019
age	Height (Krenek)	4.428	[4,304]	0.057	0.002	**0.022**	0.045
gender	Slope (Bach)	0.555	[2,304]	0.004	0.575	0.812	0.040
gender	Elbow (Bach)	2.218	[2,304]	0.015	0.111	0.295	-0.003
gender	Height (Bach)	1.002	[2,304]	0.007	0.368	0.664	-0.026
gender	Slope (Krenek)	0.191	[2,304]	0.001	0.827	0.902	-0.022
gender	Elbow (Krenek)	1.061	[2,304]	0.007	0.347	0.664	0.019
gender	Height (Krenek)	2.870	[2,304]	0.019	0.058	0.243	0.045
race	Slope (Bach)	2.312	[3,304]	0.023	0.076	0.256	0.040
race	Elbow (Bach)	0.904	[3,304]	0.009	0.439	0.664	-0.003
race	Height (Bach)	0.898	[3,304]	0.009	0.443	0.664	-0.026
race	Slope (Krenek)	0.122	[3,304]	0.001	0.947	0.980	-0.022
race	Elbow (Krenek)	0.401	[3,304]	0.004	0.752	0.867	0.019
race	Height (Krenek)	0.945	[3,304]	0.010	0.419	0.664	0.045
ExpN	Slope (Bach)	0.460	[2,304]	0.003	0.632	0.843	0.040
ExpN	Elbow (Bach)	2.481	[2,304]	0.017	0.085	0.256	-0.003
ExpN	Height (Bach)	0.020	[2,304]	0.000	0.980	0.980	-0.026
ExpN	Slope (Krenek)	2.826	[2,304]	0.019	0.061	0.243	-0.022
ExpN	Elbow (Krenek)	6.430	[2,304]	0.042	0.002	**0.022**	0.019
ExpN	Height (Krenek)	3.446	[2,304]	0.023	0.033	0.199	0.045

S2.2 Demographic association

Associations between demographic variables and the perception of temporal structures in music were explored using data from Experiments I–III combined ($N = 323$). For six parameters (`[Bach | Krenek]` × `[Slope | Elbow | Height]`), ANOVA models (`Param ~ 1 + Age + Gender + Race + Exp`) were fitted, where `Age` is a five-level factor including `Others`, `Gender` is a three-level factor including `Others`, Race is a four-level factor including `Others`, and Exp is a three-level factor representing the experiment number. The plateau height for Krenek quilts in the 18–29 age group was lower than that in the 30–44 age group (F-test for `Age`, $P_{FDR} = 0.033$; Tukey–Kramer test, $P = 0.0389$). Additionally, the elbow point for Krenek quilts was higher for Music set A (Experiments I and II) than for Music set B (Experiment III; F-test for Stimuli, $P_{FDR} = 0.033$).

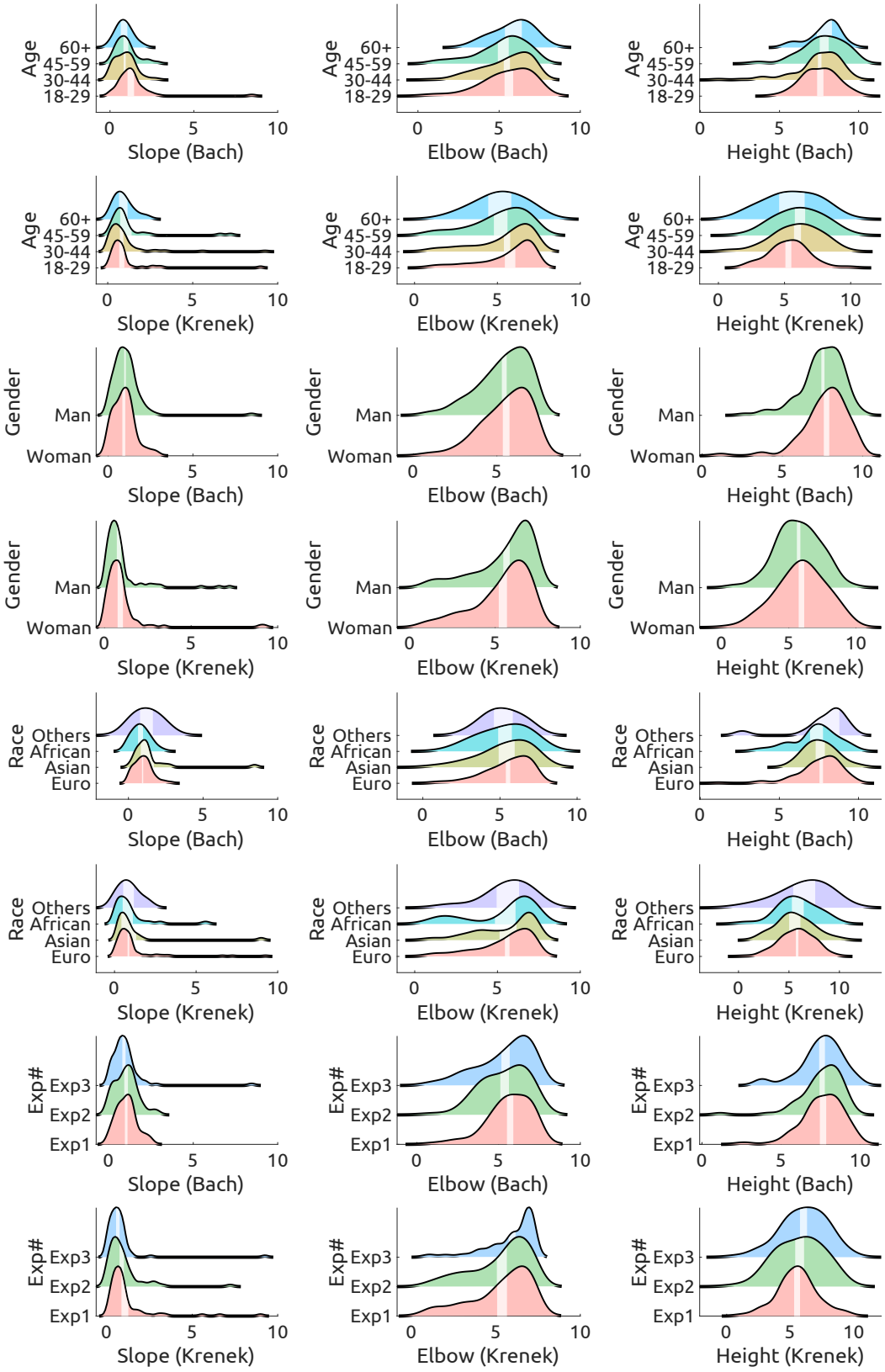

Figure S5: Group distribution of fitted parameters per level. A group with too few participants (≤ 2; gender = 'Others') was not visualised because it was difficult to estimate density functions.

Table S6: Item-wise correlation within Perceptual Abilities with age group and experiment cohort varied ($N = 304$).

Metric	Gold-MSI PA Item No.	t[297]	r[296]	P	P_{FDR}	Adj. R^2
Elbow (Bach)	1	-0.012	-0.001	0.990	0.990	0.018
Elbow (Bach)	2	1.393	0.081	0.165	0.247	0.015
Elbow (Bach)	3	2.444	0.141	0.015	$\mathbf{0.039}$	0.015
Elbow (Bach)	4	1.151	0.067	0.251	0.347	0.043
Elbow (Bach)	5	4.123	0.233	$< 10^{-4}$	$< \mathbf{10^{-3}}$	0.046
Elbow (Bach)	6	2.054	0.119	0.041	0.077	0.043
Elbow (Bach)	7	1.609	0.093	0.109	0.178	0.035
Elbow (Bach)	8	4.186	0.236	$< 10^{-4}$	$< \mathbf{10^{-3}}$	0.045
Elbow (Bach)	9	0.217	0.013	0.828	0.877	0.036
ΔHeight (Bach-Krenek)	1	0.579	0.034	0.563	0.815	0.019
ΔHeight (Bach-Krenek)	2	2.661	0.153	0.008	0.025	0.031
ΔHeight (Bach-Krenek)	3	3.074	0.176	0.002	$\mathbf{0.008}$	0.026
ΔHeight (Bach-Krenek)	4	0.502	0.029	0.616	0.815	0.04
ΔHeight (Bach-Krenek)	5	3.709	0.211	$< 10^{-3}$	$\mathbf{0.002}$	0.036
ΔHeight (Bach-Krenek)	6	2.035	0.117	0.043	0.086	0.042
ΔHeight (Bach-Krenek)	7	2.037	0.118	0.043	0.086	0.04
ΔHeight (Bach-Krenek)	8	3.068	0.176	0.002	$\mathbf{0.008}$	0.02
ΔHeight (Bach-Krenek)	9	1.07	0.062	0.285	0.514	0.039

S2.3 Item-/component-wise Gold-MSI correlation

Robust associations between the Gold-MSI subscale Perceptual Abilities and estimated parameters (Bach Elbow, Height difference) were found (Table S6)). To inform future studies, we further explored item-wise correlations. The Perceptual Abilities subscale comprises nine items (Müllensiefen et al., 2014): (1) Judge Others' Singing Abilities, (2) Recognising Novel Tune, (3) Spotting Mistakes in Performance, (4) Compare Performances, (5) Recognising Familiar Tune, (6) Judge Others' Beat Performance, (7) Judge Others' Tonal Performance, (8) Judge Own Tonal Perception and (9) Identifying Genre.

Item-wise correlations were strongest for items (3), (5) and (8) ($P_{FDR} < 0.05$). Although statistical significance was used to shorten the list, this analysis is not intended to assert any single-item relationship but rather to explore the relative relevance of individual items to temporal processing in music, thereby aiding interpretation of the observed subscale-level correlations.

In addition, principal component analysis (PCA) was used to identify orthogonal components to better understand the covariance structure across items. When correlated with the parameters, the first two components showed strong associations ($P_{FDR} < 0.05$; Table S7). Based on their PC loadings (Figure S6), the first PC reflects contributions from all items, whereas the second PC is driven primarily by

Table S7: Component-wise correlation within Perceptual Abilities with age group and experiment cohort varied ($N = 304$).

Metric	Component No.	t[297]	r[296]	P	P_{FDR}	Adj. R^2
Elbow (Bach)	1	3.133	0.179	0.002	$\mathbf{0.014}$	0.034
Elbow (Bach)	2	3.548	0.202	$< 10^-3$	$\mathbf{0.008}$	0.092
Elbow (Bach)	3	0.275	0.016	0.784	0.908	0.015
Elbow (Bach)	4	-0.091	-0.005	0.928	0.928	0.041
Elbow (Bach)	5	-0.488	-0.028	0.626	0.867	-0.008
Elbow (Bach)	6	-1.263	-0.073	0.208	0.467	0.001
Elbow (Bach)	7	-2.296	-0.132	0.022	0.08	0.008
Elbow (Bach)	8	1.067	0.062	0.287	0.574	0.002
Elbow (Bach)	9	-0.107	-0.006	0.915	0.928	-0.014
ΔHeight (Bach-Krenek)	1	3.078	0.176	0.002	$\mathbf{0.014}$	0.033
ΔHeight (Bach-Krenek)	2	2.708	0.156	0.007	$\mathbf{0.032}$	0.077
ΔHeight (Bach-Krenek)	3	1.317	0.076	0.189	0.467	0.021
ΔHeight (Bach-Krenek)	4	-0.893	-0.052	0.372	0.588	0.044
ΔHeight (Bach-Krenek)	5	0.388	0.023	0.698	0.898	-0.009
ΔHeight (Bach-Krenek)	6	0.858	0.05	0.392	0.588	-0.002
ΔHeight (Bach-Krenek)	7	-1.479	-0.086	0.14	0.421	-0.002
ΔHeight (Bach-Krenek)	8	0.95	0.055	0.343	0.588	0.001
ΔHeight (Bach-Krenek)	9	0.244	0.014	0.807	0.908	-0.014

Items (3), (5), and (8), consistent with the item-wise correlation results.

However, again, this analysis is not intended to single out specific items, but rather to aid interpretation.

References

Broze, Y., & Huron, D. (2013). Is higher music faster? pitch–speed relationships in western compositions. *Music Perception: An Interdisciplinary Journal*, *31*(1), 19–31.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS one*, *9*(2), e89642.

Norman-Haignere, S. V., & McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS biology*, *16*(12), e2005127.

Overath, T., McDermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature neuroscience*, *18*(6), 903–911.

Pasler, J. (1985). Ernst křenek-in retrospect. *Perspectives of New Music*, 424–432.

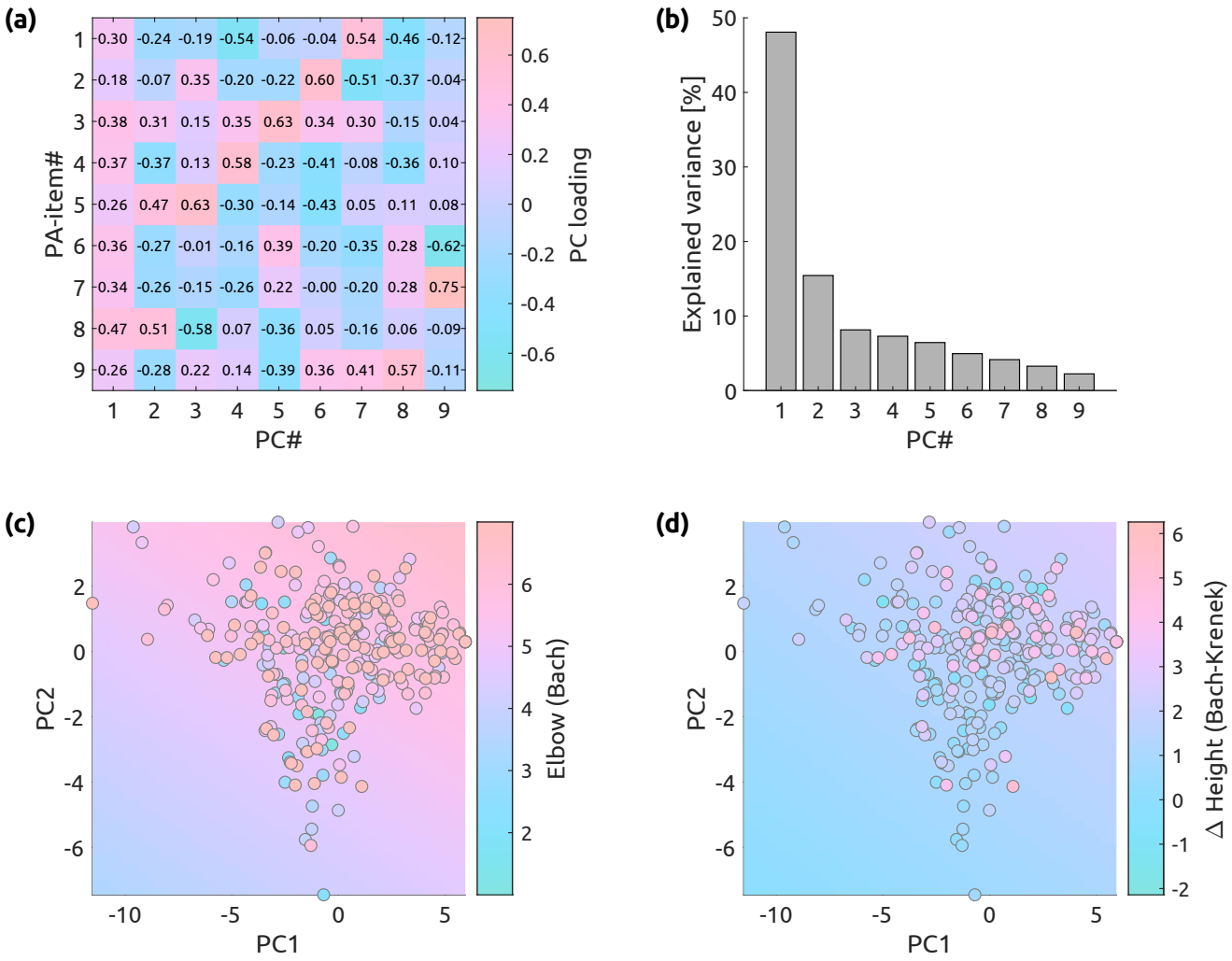

Figure S6: Visualisation of principal component analysis. (a) Loadings of items on components. (b) Explained variance per component. (c–d) scatter plots of PC scores (i.e., eigenvariates) of the first two components over fitted planes with the respective elbow-function parameters as response variables (c, Elbow for Bach; d, Height difference). Each marker indicates a participant, with the respective parameters colour-coded. The colour of the planes indicates fitted values.