The Hierarchical Structure of Temporal Modulations in Music is Universal across Genres and matches Infant-Directed Speech.

Tatsuya Daikoku [1,2,] *, Usha Goswami [1]

[1] Centre for Neuroscience in Education, University of Cambridge, Cambridge CB2 3EB, United Kingdom

[2] International Research Center for Neurointelligence, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

Correspondence: Tatsuya Daikoku

Email: daikoku.tatsuya@mail.u-tokyo.ac.jp

Abstract

Rhythm is essential in music. It is formed by the hierarchical structure of events such as notes, beat, accent and grouping. The core acoustic correlates of these hierarchical rhythmic events, however, remain unknown. This study examined the slow (<40 Hz) temporal modulation structure of music by applying two modeling approaches previously applied to infant- and child-directed speech. The first approach utilized a low-dimensional representation of the auditory signal based on the cochlear filterbank in the human brain, comprising 3 broad bands of amplitude modulation and 5 broad spectral bands; a "Spectral-Amplitude Modulation Phase Hierarchy (S-AMPH) modelling approach. The second utilized probabilistic amplitude demodulation (PAD), a signal-driven modelling approach that makes no adjustments for the brain. PAD has produced successful models of natural sounds like wind and rain, which are characterized by amplitude modulation 'cascades' correlated over long time scales and across multiple frequency bands. When applied to music, both models revealed a very similar hierarchically-nested amplitude modulation spectrum across different musical genres and instruments, including song. Accordingly, the core temporal architecture yielding musical rhythm appears universal. This same architecture is revealed by modelling infant- and child-directed speech, suggesting that both music and language may depend on a shared domain-general amplitude modulation architecture which, early in life, relies on shared neural processes. The demonstration that temporal modulation bands play a key role in rhythm hierarchies in music as well as in speech could suggest that the same evolutionary adaptations may underpin both music and language, explaining why both are human universals.

Keywords

Amplitude modulation hierarchy; Rhythm; Music; Oscillation; Phase synchronisation

Significance Statement

- Novel modelling across musical genres reveals a universal hierarchy of temporal modulations underpinning musical rhythm.
- The temporal modulation architecture of music shares fundamental commonalities with infant- and child-directed speech (IDS, CDS).
- As in IDS and CDS, temporal rhythm in delta- and theta-rate amplitude modulation bands shows mutual phase dependence.

1. Introduction

1.1. Hierarchical Modulation Structure of Auditory Rhythm

Music is a universal language that is created and enjoyed by people from vastly different cultures and ages, appearing in all societies so far measured [1]. Rhythm is a fundamental component of music, and also an essential aspect of all natural languages [2]. Speech rhythm, including intonation and accentuation, is indispensable for infants learning the phonological structure of human languages, for understanding semantics, and can also convey the speaker's emotion and communicative intent [3,4]. As rhythm is ubiquitous in both music and language, it is plausible to propose that evolution played a common role in configuring their acoustic structure, also allowing them to interact with each other as in human song.

It is well known that the speech signal can be represented as a nested hierarchy of phonological units including prosody, syllable, rhyme and phoneme [5,6]. Recent evidence from adults suggests that these linguistic units can be identified from patterns of amplitude modulation (AM) in the acoustic sound waveforms of speech [7]. Meanwhile, related modelling of infant- and child-directed speech (IDS and CDS) has shown that the temporal patterning of prosodic, syllabic, rhyme and phoneme linguistic levels are yielded by the phase relations of temporal modulation bands in the amplitude envelope which correspond temporally to delta, theta and beta/gamma EEG bands, respectively [8,9]. An amplitude modulation phase hierarchy modelling approach has thus produced clear evidence regarding the acoustic foundations of the rhythm structure of IDS and CDS. Nevertheless, it remains unclear how musical rhythm could be hierarchically represented in the temporal envelope of musical waveforms.

Music shares many commonalities with speech at both an acoustic and cognitive level [10–12]. At the acoustic level, both utilize spectral (e.g., pitch and timbre) and temporal (e.g., rhythm) cues to convey information [13]. At a cognitive level, processing both music and speech requires similar learning, memory, and attention systems [4]. The listener needs to integrate discrete acoustic events into a coherent perceptual stream with accentuation and rhythms, and even infants appear able to do this [14]. The perception of rhythmic timing in both music and speech has been linked to 'P-centres', the perceptual moment of occurrence ('perceptual centre') of each beat or syllable for the listener [15,16]. Related research has revealed that amplitude rise times are critical to this perceptual experience (amplitude rise times are typically called attack times in the musical literature, [17,18]). In addition, perceiving the beat structure in music depends on recognizing that some notes have greater accentuation than others, akin to perceiving strong and weak (stressed and

unstressed) syllables in speech. In children, recognizing accentuation and beat in music and perceiving strong and weak syllable patterns in speech are known to be developmentally inter-related, with both abilities also related to children's amplitude rise time discrimination [19]. Accordingly musical rhythm, like the phonological hierarchy yielded by speech rhythm, may also comprise a nested hierarchy of several units such as accents, beats, and various lengths of notes (half note, quarter note, etc.), and may also depend on patterns of amplitude modulation in the acoustic sound waveform cued by different amplitude rise times.

Linguistically, hierarchical structures are classically represented as a tree or grid that captures the *relative prominence* of phonological units [2,6,20]. In the tree representation, a 'parent' node (element) at one tier of the hierarchy encompasses one or more 'daughter' nodes at a lower level of the hierarchy. The adjacent connection between the parent and daughter nodes are indicated as 'branches' in the tree. So a parent node such as the trisyllabic word "pussycat", which is also the prosodic foot, would have 3 daughter nodes at the next hierarchical level, comprising the three syllables. It has long been recognised that such tree representations may also model the core principles of metrical structure in music, and that tree representations of hierarchical rhythmic structure could be shared between speech and music [21]. In music, the term 'meter' refers to the number of beats per bar (e.g. 2/4, two quarter-notes per bar). If the musical beats and the bar are analogous to syllables and prosodic feet, then meter will equate to the number of syllables per prosodic foot. The acoustic underpinnings of this hierarchical rhythmic structure of speech in terms of syllables per prosodic foot has been modelled using nursery rhymes, one common form of CDS [8]. English nursery rhymes like "Jack and Jill went up the hill" and "Pussycat pussycat where have you been" provide examples of 2 (trochaic rhythm) versus 3 (dactyl rhythm) syllables per foot. Leong and Goswami modelled the AM structure of this form of CDS and revealed an AM phase hierarchy that was consistent across the spectral range of speech (modelled as 5 bands) comprising 3 temporal bands of AMs, equating temporally to electrophysiological rhythms found across the brain (at the oscillatory rates of delta, theta and beta-low gamma, see also [22]). To date, no studies have examined the temporal correlates of the musical rhythm hierarchy from the same modelling perspective. We remedy that omission here, exploring whether a universal hierarchical structure of musical rhythm will be revealed using a similar modelling approach across different musical genres and instruments.

1.2. Biological Underpinning: Neural Oscillators

A primary acoustic correlate of perceived rhythmic patterns is the amplitude envelope of "*temporal modulation*" below about 40 Hz, which reflects fluctuations of signal energy (i.e., loudness or sound intensity) related to onset timing, accent and chunking. The temporal modulation spectrum of sounds has a relatively straightforward neurophysiological interpretation. It is well known that the cochlea can decompose acoustic signals into narrow frequency bands [23]. These narrowband signals are then half-wave rectified and low-pass filtered at different stages of the auditory pathway [24], thereby extracting the temporal modulation envelope of the narrowband signals [25]. In auditory cortex, oscillatory neural networks track the temporal modulation envelope below about 40 Hz [26,27], with different oscillators (delta, theta, gamma) phase-synchronizing with different amplitude modulation (AM) patterns at matching rates, using amplitude rise times as automatic triggers for phase re-setting [28–30]. These populations of auditory neurons thereby encode modulation spectra of the temporal envelope at different neural signalling rates [31–39]. This temporal information contributes to building perceived rhythmic structures, while the spectral information in speech constitutes pitch and other acoustic features including noise.

This neural tracking of the temporal modulation structure of speech contributes to parsing of the signal into several hierarchical chunks, enabling listeners to recognize phonological units such as syllables and words [40], and also musical units such as such as notes, beat, and meter [41]. The populations of cortical neurons are known to fluctuate in excitability at different frequency rates such as delta (<4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–80 Hz). These oscillatory bands play important roles in different levels of phonological perception [34,42,43]. For example, delta, theta, and beta/gamma oscillators in the auditory cortex contribute to the perception of prosodic, syllabic, and phonetic information respectively [34,44,45]. Regarding acoustic processing, there is also growing biological evidence that these adjacent-band oscillators are not independent of, but interdependent on, each other [46,47]. For example, the phase of delta oscillators modulates the phase of theta oscillators, and theta phase modulates beta/gamma power. This cascaded oscillatory system is thought to contribute to encoding the hierarchy of acoustic modulations, which in turn contribute to the parsing of large (e.g., prosodic phase) and smaller linguistic units (e.g., syllables)[46].

Given that the temporal modulation spectrum of sounds matches the spectrum of neuronal oscillations, it may follow that musical rhythm will induce phase-locking and synchronisation of neural oscillators to temporal modulation patterns in the same way as speech rhythm. Such phase locking may yield the perceptual experience of musical components such as minim, crotchet, and quaver as well as beat, tempo, and musical phrasing. The present study, as a first step in exploring

this possibility, investigates whether there are any consistencies in the hierarchical structure of temporal modulation spectra of music across a range of genres and instruments.

1.3. Purpose of This Study

The aim of the present study is to explore the modulation architecture of natural music across genres and instruments, including the human voice. The music samples consisted of over 39 h of recordings (sampling rate = 44.1 kHz) of Western music across a sub-set of genres (Western-classical music, Jazz, Rock, and Children's song) and musical instruments (single-voice: Violin, Viola, Cello, and Bass; multi-voice: Piano and Guitar) (see S1 Appendix). The modulation spectra were analysed using two different algorithms, the Spectral-Amplitude Modulation Phase Hierarchy (S-AMPH, [8]), and Probabilistic Amplitude Demodulation (PAD, [48]). Both models enable music signals to be parsed into an hierarchy of temporal modulation structures, but the principles underpinning their approaches are different. The S-AMPH simulates the frequency decomposition by the cochlea [23,49], while PAD models the statistical structure of natural sounds, using probabilistic amplitude demodulation with no adjustments for the brain (see STAR Methods). Given the biological evidence that each oscillator modulates the adjacent-band oscillator in speech perception [46,47], we also examine mutual dependency between tiers of the resulting temporal hierarchies. We hypothesize that, as suggested by the metrical linguistics of tree representations, the adjacent tiers will be highly dependent on each other compared with non-adjacent tiers. The present study is the first to explore whether a mutually-dependent hierarchical tree-based structure of AMs underlies the perceptual experience of musical rhythm. On the assumption of universality of an AM hierarchy in music as well as speech, it is hypothesized that acoustically-emergent musical units can be parsed from the resulting temporal modulation architectures via structure-mapping. It is also hypothesized that the two models will show consistent findings regarding hierarchical AM patterns, with both models demonstrating the same core principles underlying the structure of musical rhythm. If so, this would constitute evidence that the hierarchy of the temporal modulation architecture in music is universal, regardless of the type of model applied (i.e., filtering vs. probabilistic).

2. Results

2.1. Signal Processing

The S-AMPH model creates a low-dimensional representation of the auditory signal, using equivalent rectangular bandwidth (ERB_N) filterbanks [23,49]. This generates an hierarchical representation of the core spectral (acoustic frequency spanning 100–7,250 Hz) and temporal (oscillatory rate spanning 0.9–40 Hz) modulation architectures in the auditory envelope (see S2 Appendix). The number and the edge of bands are determined by principal component analysis (PCA) dimensionality reduction of original high-dimensional spectral and temporal envelope representations (Fig 1). This modulation filterbank can generate a cascade of amplitude modulators at different oscillatory rates, producing the AM hierarchy. However, the filterbank may partially introduce artificial modulations into the stimuli because the bandpass filters can introduce modulations near the center-frequency of the filter through “ringing”. Therefore, we also used a second AM-hierarchy extraction method (i.e., PAD) as a control.

The PAD model does not implement the Hilbert transform, filtering, and PCA, but infers the modulators and a carrier based on Bayesian inference, identifying the envelope which best matches the data and the *a priori* assumptions. PAD can be run recursively using different demodulation parameters each time, generating a cascade of amplitude modulators at different oscillatory rates (i.e., corresponding to neural rates of delta, theta, alpha, beta, and gamma), forming an AM hierarchy [50] (Fig 2).

2.2. Amplitude Modulation Properties of Music from S-AMPH

To determine the number and the edge of the core spectral (acoustic frequency spanning 100–7,250 Hz) and temporal (oscillatory rate spanning 0.9–40 Hz) modulation bands, PCA was applied separately for spectral and temporal dimensionality reductions (for more detail, see STAR Methods). The PCA has been used for dimensionality reduction in speech studies (e.g. [51,52]). The present study focused on not the component scores, but the absolute value of component loadings, which indicate the underlying patterns of correlation between the high-dimensional channels. That is, the PCA loading was adopted to identify patterns of covariation between the high-dimensional channels of spectral (28 channels) and temporal (24 channels) modulations, and to determine groups (or

clusters) of channels that belong to the same core modulation bands. In spectral and temporal PCAs, the 28 and 24 spectral channels were taken as separate variables (Fig 1 and S2 Appendix), yielding a total of 28 and 24 principal components, respectively. On the other hand, only the first 5 and 3 components were considered in spectral and temporal PCAs respectively, because these already cumulatively accounted for over 55% (on average) of the total variance in the original sound signal. S3 Appendix shows loading patterns in each music genre for the spectral and temporal PCA. Then, peaks and troughs in the grand average PC loading patterns were identified because they reflect centre and boundaries of co-modulated clusters of channels. A minimum peak-to-peak distance of 2 and 5 channels was set for the spectral and temporal PCAs, respectively. The core spectral and temporal modulation bands were determined based on the criteria that at least 2 of the 5 PCs showed a peak, and that at least 1 of the 3 PCs showed a peak for at least 2 out of the 5 spectral bands, respectively.

Spectral PCA

S3 Appendix shows the loading patterns and cumulative contribution ratios in each music genre and the grand average. The first to fifth principal component (PC1 to PC5) accounted for, on average, 33%, 8%, 6%, 5%, and 5% of the total variances, respectively. One peak (~3000 Hz) was identified from the loading pattern of PC1. Thus, PC1 was assumed to reflect the global correlation between spectral channels. The peak of ~300 Hz and the "*flanking*" trough of ~350 Hz are identical between PC3 and PC4, providing corroborating evidence for a lowest spectral band at this spectral location with a potential boundary between the first and second spectral bands at ~350 Hz (troughs indicate potential boundaries between modulation rate bands). A close peak of around 420-600 Hz and the identical flanking trough of ~700 Hz in PC3, PC4 and PC5, providing corroborating evidence for a second spectral band at this spectral location and a potential boundary between the second and third spectral bands at ~700 Hz. A close peak of 900-1100 Hz n PC3, PC4 and PC5, and a close flanking trough of 1750-2000 Hz in PC2, PC3 and PC5, provide corroborating evidence for a third spectral band at this spectral location and a potential boundary between the third and fourth spectral bands at ~1750 Hz. The peak of ~2500 Hz and the flanking trough of ~3900 Hz are identical between PC3 and PC5, providing corroborating evidence for a fourth spectral band at this spectral location and a potential boundary between the fourth and fifth spectral bands at ~3900 Hz. A peak of ~5500 Hz is identical between PC2 and PC3, providing corroborating evidence for a fifth spectral band at the spectral location. Further, this fifth peak is the highest peak observed in the spectral PCA loading patterns. Based on the criteria (see STAR Methods), the spectral PCA provided evidence for the presence of 5 core spectral bands in the spectral modulation data (300, 500, 1000, 2500 and 5500 Hz), with at least 2 out of 5 PCs showing peaks in each of these 5

spectral regions. Furthermore, we consistently observed 4 boundaries between these 5 spectral bands (350, 700, 1750 and 3900 Hz). Table (a) in the S3 Appendix provides a summary of these 5 spectral bands and their boundaries.

It is noteworthy that these 5 spectral bands are proportionately-scaled with respect to the logarithmic frequency sensitivity of human hearing. Additionally, these results are similar to the spectral bands modelled in CDS [8]. This is suggestive of universal characteristics regarding the spectral modulation features of speech and music. The present study, however, considered the core spectral bands using the grand average PC loading patterns across music genres. It is reasonable to expect that there will be some individual variation in peak and trough location across genres. This individual variation is explored and described further in S3 Appendix. It can be noted that the loading patterns for the 5 PCA components showed roughly similar characteristics across the genres. We, however, also find some individual variation at each spectral modulation band.

Temporal PCA

S3 Appendix shows the loading patterns and cumulative contribution ratios in each music genre and Fig 3 shows the grand average loading patterns (absolute value) for the first three principal components arising from the temporal PCA of each of the 5 spectral bands determined in the spectral PCA (Table a in Appendix 3). The colours in Fig 3 represent the 5 spectral bands, while the types of lines (i.e., bold, dashed and dotted) represent PCs 1,2 and 3, respectively. As may be clearly observed, the loading showed consistent patterns between the 3 PCA loading patterns.

The first to third principal component (PC1 to PC3) accounted for, on average, 49%, 11% and 6% of the total variances, respectively (for more detail, see Table b in S3 Appendix). PC1 showed a moderate peak at acoustic frequencies of 7-9 Hz in all of the 5 spectral bands. As observed in the spectral PCA, PC1 in the temporal PCA might reflect the global correlation between temporal channels. As no troughs were detected in PC1 (indicating no potential boundaries), our analysis focused on PC2 and PC3. The loading patterns of PC2 result in 2 strong peaks at acoustic frequencies of 1-2 Hz (evidence for a delta-rate band of AMs) and 20-30 Hz (evidence for a beta-gamma rate band of AMs), and 1 strong flanking trough at acoustic frequencies of ~7 Hz. These findings are consistent between the 5 spectral bands, suggesting the potential existence of at least 2 core temporal bands. Compared with PC1 and PC2, PC3 loading patterns varied across spectral bands. As detected in PC2, all the spectral bands showed a peak in loading at ~30 Hz, and spectral band 2 similarly showed a peak at ~1 Hz. PC3 also showed an additional mid-rate peak at ~5 Hz

(theta-rate band) and ~9 Hz (alpha-rate band). The flanking troughs for these peaks occurred 2-3 Hz and 15-18 Hz.

Based on the criteria (Methods), the temporal PCA provided evidence for the presence of 4 core bands, with at least 1 out of 3 PCs showing peaks in each of these 4 temporal regions. Furthermore, we consistently observed 3 boundaries between these 4 temporal bands. Table 1 provides a summary of these 4 temporal bands and the boundaries. It is important to note that, as shown in the spectral PCA, the results of the temporal PCA were highly similar to studies of CDS by Leong and Goswami [8] and to related studies of IDS [9]. Further, overall frequencies showed higher consistency across genres. These findings are suggestive of universal characteristics of temporal modulation structures in speech and music.

Summary of S-AMPH results

Using an ERB_N filterbank that simulates the frequency decomposition by the cochlea [23,49], the S-AMPH model generated an hierarchical representation of the core spectral (acoustic frequency spanning 100–7,250 Hz) and temporal (oscillatory rate spanning 0.9–40 Hz) modulation architectures in the envelope of a range of musical genres. In each genre, the AM hierarchies consisted of 5 spectral bands (100–350 Hz; 350–700 Hz; 700–1750 Hz; 1750–3900 Hz; 3900–7250 Hz), each containing 4 hierarchically-nested temporal bands ([1] 0.9–2.5 Hz, [2] 2.5–7 Hz, [3] 7–17 Hz, [3] 17–40 Hz).

2.3. Amplitude Modulation Properties of Music from PAD

The PAD model implements amplitude demodulation by estimating the most appropriate modulator (envelope) and carrier based on Bayesian inference that identifies the envelopes which best match the data and the *a priori* assumptions (for more detail, see STAR Methods). PAD can be run recursively using different demodulation parameters each time, generating a cascade of amplitude modulators at different modulation rates, forming an AM hierarchy [50] (Fig 2). The present study manually entered the PAD parameters, producing modulators that could correspond to each of five neural oscillatory rate bands (i.e., delta: 0-4 Hz, theta: 4-8 Hz, alpha: 8-12 Hz, beta: 12-30 Hz, and gamma: 30 Hz+). Note that manual entry of these parameters does not predetermine the results,

rather it enables exploration of whether there is a prominent peak frequency observed in each oscillatory rate band, and which such frequencies may comprise the core temporal modulation architecture. The carrier was interpreted as components including noise and pitches whose frequencies are much higher than the core modulation bands in the phrasal, prosodic, syllabic, and phonetic components. In each of the music samples, the modulators (envelopes) of the five oscillatory bands were converted into frequency domains by FFT. We hypothesized that modelling with PAD should reveal the same core principles as the S-AMPH modelling regarding hierarchical AM patterns.

S3 Appendix shows the modulation spectra of FFT in each music genre. It is reasonable to expect that there will be some individual variation in peak and trough location across genres. This individual variation is explored and described further in S3 Appendix. Overall, low spectral frequencies (~5Hz) showed higher consistency across genres while mid-to-high spectral frequencies showed some individual variation. It is of note that these tendencies are similar to the findings regarding spectral modulation in the S-AMPH modelling, providing corroborating evidence for consistency of outcome between the different modelling approaches. Fig 3 represents the grand average modulation spectra of FFT for each of oscillatory band. As can be seen, four peak frequencies were detected at ~2.4 Hz, ~4.8 Hz, ~9 Hz and 16 Hz. Accordingly, PAD showed similar findings to the S-AMPH model: both S-AMPH and PAD detected peak frequencies in AM bands corresponding temporally to neural delta, theta, alpha and beta/gamma neural oscillatory bands (see Table 1).

2.4. Mutual Information in Both Models

We also examined mutual dependency between tiers of the temporal hierarchy using mutual information analyses. Mutual information is a measure of the mutual dependence between two variables. Hence, this analysis enables computation of whether a certain AM rhythm (i.e., AM rhythms corresponding to delta, theta, alpha and beta neural oscillations) is dependent on another AM rhythm. Previous evidence from speech and rhythmic sounds suggests that adjacent-band neural oscillators are not independent of, but interdependent on, each other [46,47]. Accordingly, we hypothesized that adjacent AM tiers would connect via "*branches*" in the AM hierarchy and would be mutually dependent on each other compared with non-adjacent AM tiers. If so, this would

represent evidence for an hierarchical "*tree-based*" structure of musical rhythm, highlighting the universality of the AM hierarchy in music as well as speech.

The MI analysis adopted the phase angle "*θ*" of the core temporal modulation envelopes corresponding to neural delta, theta, alpha and beta/gamma waves, which was detected in each of the S-AMPH and PAD approaches. That is, in the S-AMPH model, the 5 spectral envelopes (see Fig 1b) were passed through a second series of band-pass filters to isolate the 4 different AM bands based on the results of the temporal PCA (channel edge frequencies: 0.9, 2.5, 7, 17 and 30 Hz). The phase angles were then calculated using each of the 4x5 temporal modulation envelopes. In PAD, the phase angles were calculated using the four core modulators (envelopes) that had been detected in the previous analyses, which correspond to delta, theta, alpha, and beta/gamma neural bands, respectively (for example, k, l, g, e in Fig 2). Then, using the phase angle values, the mutual information between different temporal modulation bands was measured.

S4 Appendix shows the PSI in each music genre and the grand average. We detected high consistency between music genres in both S-AMPH and PAD models. Therefore, to understand a universal property of music, we focused on the grand average (Fig 4). The results showed that adjacent tiers of the AM hierarchy are mutually dependent on each other compared with nonadjacent tiers. Further, mutual dependence between delta- and theta-rate AM bands was the strongest of all mutual dependence in both models. This highlights another similarity to the speech-based modelling. When adults were asked to identify tone-vocoded nursery rhymes, a procedure resulting in acoustic patterns with clear rhythmic temporal patterning but that sounded like morse code or flutter, they relied solely on the phase relations between the AMs in the two slowest AM bands, corresponding to the neural oscillatory delta and theta bands [7]. For speech, Leong et al. argued that this novel acoustic statistic ("delta-theta AM phase alignment") could be used to identify the rhythm pattern of a phrase, which for speech means its prosodic structure. The current modelling suggests that for music, delta-theta phase alignment of AM bands may underpin metrical structure.

2.5. Phase Synchronization in both Models

We next investigated the integer ratios between "*adjacent*" AM hierarchies (i.e., the number of parent vs. daughter elements in an AM hierarchy). This analysis aimed to identify how many

daughter elements a parent element could encompass in general. We adopted the core temporal modulation envelopes corresponding to the neural rates of delta, theta, alpha and beta/gamma waves, which was detected in each of the S-AMPH and PAD approaches. As in the mutual information analysis, this analysis made use of the 4x5 temporal modulation envelopes of the S-AMPH model and the 4 temporal modulation envelopes of the PAD model. A Phase Synchronization Index (PSI) was then computed between adjacent modulation rate bands in the S-AMPH representation for each of the five spectral bands and in the PAD representation (i.e., delta vs. theta, theta vs. alpha, alpha vs. beta, beta vs. gamma phase synchronizations). The n:m PSI was originally conceptualized to quantify phase synchronization between two oscillators of different frequencies (e.g., muscle activity; [53]), and was subsequently adapted for neural analyses of oscillatory phase-locking [54]. For example, if the integer ratio is 1:2, then the parent element encompasses 2 daughter elements for the rhythm.

S5 Appendix shows PSI in each music genre and the grand average. We detected high consistency between music genres in both S-AMPH and PAD models. Therefore, we again focused on the grand average (shown in Fig 5). The PSI of the S-AMPH model suggested that the PSI of 1:2 integer ratios is the highest in all of the adjacent oscillatory bands. The PSIs of 1:3 and 2:3 integer ratios are also higher than the other integer ratios, suggesting that the simpler integer ratios (i.e., m/n) are likely to synchronize in phase with each other. The PSIs of the PAD model were similar to the S-AMPH, but PAD was more sensitive to the simple integer ratios. In PAD, the PSIs of not only the 1:2 integer ratios, but also those of the 2:3, 3:4 and 4:5 integer ratios were notably higher than the other integer ratios. This may arise because the filterbank used in the S-AMPH model may partially introduce some artificial modulations into the stimuli through “ringing”. However, the ERB_N filterbank in the S-AMPH model is the filtering process that reflects the frequency decomposition by cochlear function in the normal human ear. Hence, the different findings between S-AMPH and PAD models regarding phase synchronization may imply that there are differences between auditory signals as perceived by the human brain and the purely physical and statistical structure of music.

Nevertheless, as shown in Fig 5, the PSI between delta- and theta-rate AM bands was consistently the largest PSI in both the S-AMPH and PAD models. As described earlier (section 2.4 and Fig 4), this consistency was also detected in the mutual information analyses, as the mutual dependence between delta- and theta-rate AM bands was the strongest of all mutual dependencies. As noted, prior listening data from adults showed that a strong syllable was perceived when delta and theta modulation peaks were in alignment[7]. Accordingly, the strong rhythmic character and acoustic temporal regularity of both speech and music appears to be influenced by amplitude modulations

in the delta band. The present findings indicate that music has similar properties to IDS and CDS: phase alignment between delta- and theta-rate AM bands contributes to establishing metrical musical rhythm patterns as well as to establishing strong versus weak syllable patterning in speech. Indeed, speech-based modelling has revealed that IDS shows significant extra modulation energy in the delta band compared to adult-directed speech, contributing to its high rhythmicity [9]. The current data also map nicely to the Plomp and Levelt [55] modeling of the dissonance curve for spectral frequency (i.e., "*pitch*"). Our modelling data for temporal frequency (i.e., "*rhythm*") show a similar structure to their data (shown in Fig 5c). This may imply that these physical properties of fast spectral frequencies are also involved in very slow temporal modulation envelopes below 40 Hz.

2.6. Simulation Analyses

To investigate whether the detected (dissonance curve-like) characteristics really represent universal features of natural musical rhythm, we conducted simulation analyses. We generated synthesized sounds that consisted of four temporal modulation envelopes (i.e., modulator) and one spectral frequency (carrier). That is, 2 Hz, 4 Hz, 8 Hz and 16 Hz sine waves were summarized to synthesize one compound tone waveform. The compound tone waveform was then multiplied by 200 Hz sine waves. The synthesized waveform comprised a sound that included temporal information of delta, theta, alpha and gamma rhythms, and spectral information of a pitch around that of natural human voices (for the figure, see S5 Appendix). It is important to note that all of the temporal envelopes were comprised of simple sine waves with frequencies of a power of 2. Hence, we can hypothesize that 1:2 integer ratios should clearly and consistently appear compared with other integer ratios. However, the analyses showed that the PSIs of natural music were different from those for artificial sounds. These data show that natural musical rhythm has covert universal integer ratios within the AM hierarchy.

3. Discussion

The modelling presented here has revealed a universal hierarchy of temporal architecture underpinning the perception of musical rhythm. Using two different modelling approaches, it has been shown that acoustically-emergent musical units can be parsed from the temporal modulation spectra of all musical genres, and are reflected in each of delta-, theta-, alpha- and beta/gamma-rate bands of amplitude modulation (Fig 6). This modulation architecture matches that found in comparable modelling studies of both IDS and CDS (language spoken to infants aged either 7, 9 or 11 months, and nursery rhymes; [8,9]). As the brain begins learning language using IDS, and consolidates this learning via the rhythmic routines of the nursery (CDS), the present findings suggest that both music and language rely on the same neural processes. Both music and language may depend on a shared domain-general amplitude modulation architecture which yields the perceptual experience of rhythm, and is encoded by cortical oscillations. Prior claims that the rhythmic properties of music and language are distinct [56], with the modulation spectrum for music peaking at 2 Hz but peaking at 5 Hz for speech, appear to arise from the exclusive reliance of the speech modelling on ADS. By contrast, our modelling approach was derived from studies of IDS and CDS, and suggests that the statistical regularities in temporal modulations are the same for music, IDS and CDS. The demonstration that temporal modulation bands play a key role in rhythm hierarchies in music as well as in speech may also suggest that the same evolutionary adaptations underpin both music and language. In turn, this may explain why both music and language are human universals. To the best of our knowledge, our study is the first to reveal a universal hierarchical structure of musical rhythm across difference genres and instruments.

The data also revealed strong mutual dependence between adjacent bands in the AM hierarchy across musical genres, for both modelling approaches. In particular, the mutual dependence between delta- and theta-rate bands of AM was the strongest dependence identified by both models. In addition, the results from the phase synchronisation analyses showed that PSIs between delta- and theta-rate AM bands were higher than the other PSIs regardless of the integer ratios. The phase alignment of delta- and theta-rate bands of AM has been suggested to be a key acoustic statistic for the language-learning brain, with individual differences in sensitivity to these AM rates and their associated rise times implicated in disorders of language development [57]. In speech-based studies, the phase alignment between the slower AM bands (delta and theta) has been demonstrated to play a key role in the perception of metrical patterning (judging whether tone-vocoded nursery rhymes were trochaic or iambic) [7]. Perceptual data from adults showed that a strong syllable was perceived when delta and theta modulation peaks were in alignment. The present findings concerning mutual dependence and phase synchronisation indicate that music

also has such properties: phase alignment between delta- and theta-rate AM bands contributes to establishing musical metrical structure as well.

Another interesting result from the phase synchronisation analyses was the appearance of universal integer ratios within the AM hierarchy. While the 1:2 integer ratio was strongest for both models, the PSIs for 1:3 and 2:3 were also higher than the other integer ratios explored, for both models. For the PAD modelling approach, which does not make any adjustments for the cochlea, the 2:3, 3:4 and 4:5 integer ratios were also prominent. This statistical patterning may be related to the different metrical structures and integer ratios that characterise music from different cultures [1,58]. Our modelling suggests that temporal modulations in music (the different AM bands) are related by integer ratios in a similar way to the integer ratios relating notes of different fundamental frequencies that create harmonicity (see the similarity between the PSIs for the two models shown in Figure 5 and the dissonance curve measured by Plomp & Levelt [55]). Converging prior modelling of speech has shown that the probability distribution of amplitude–frequency combinations in human speech sounds relates statistically to the harmonicity patterns that comprise musical universals [59]. Our modelling approach suggests that amplitude modulation phase hierarchies may play as strong a role as harmonicity regarding universal mechanisms of hearing that are important for both music and language. Further research focused on the integer ratios foregrounded by both amplitude modulation phase hierarchy approaches and harmonicity approaches may illuminate these perceptual universals further.

At the applied level, the possible utility of musical interventions for children with disorders of language learning such as developmental language disorder (DLD) and developmental dyslexia is supported by our modelling [60]. Such interventions are likely to be most beneficial when the underlying hierarchy of the temporal structure of music corresponds to the temporal hierarchy underpinning speech rhythm [57]. The evidence presented here regarding the perceptual commonalities between musical and speech rhythms implies that careful matching of the statistical rhythm structures across music and speech for different languages would lead to better remedial outcomes [61,62]. Similar interventions could be beneficial for second language learners. Music is a universal language. However, to date only one stress-timed language, English, and one syllable-timed language, Portuguese, have been modelled using the S-AMPH and PAD approaches [7–9,63]. Accordingly, further studies are necessary to understand how music interventions can contribute to improving speech processing in different languages. Nevertheless, the current findings may provide new interdisciplinary perspectives for both musical pedagogy and speech and language therapies, related to both temporal processing and to rhythmic movement [64,65].

In conclusion, the present study revealed a universal hierarchy of core temporal architecture in musical rhythm. The temporal modulation properties of music were highly similar whether modelled using the S-AMPH or PAD, with AM bands playing a key role in organising the musical hierarchy. The present study is the first to demonstrate that acoustically-emergent musical units can be parsed from the temporal modulation architecture of different musical genres via structure-mapping of the mutually-dependent phase relations within this AM hierarchy. The data suggest that universal perceptual mechanisms based on AM hierarchies underpin both language and music, and that these universal perceptual mechanisms can also explain cross-cultural regularities in musical systems [58].

4. Materials and Methods

The music samples consisted of over 39 h of recordings (sampling rate = 44.1 kHz) of Western music across genres (Western-classical music, Jazz, Rock, and Children's song) and musical instruments (single-voice: Violin, Viola, Cello, and Bass; multi-voice: Piano and Guitar). This corresponds to music corpora used in the study by Ding et al. [56], but we also included the children's song to characterize a more general property of modulation spectra across a variety of music genres. The full list is provided in S1 Appendix. The acoustic signals were normalized based on z-score (mean = 0, SD = 1). The spectro-temporal modulation of the signals was analysed using two different algorithms for deriving the dominant amplitude modulation (AM) patterns: [1] Spectral Amplitude Modulation Phase Hierarchy (S-AMPH, [8]), and [2] Probability Amplitude Demodulation based on Bayesian inference (PAD, [48]). The S-AMPH model is a low-dimensional representation of the auditory signal, using equivalent rectangular bandwidth (ERB_N) filterbank, which simulates the frequency decomposition by the cochlea function in a normal human [23,49]. This model can generate a hierarchical representation of the core spectral (acoustic frequency spanning 100–7,250 Hz) and temporal (oscillatory rate spanning 0.9–40 Hz) modulation architectures in the auditory envelope such as speech and music. This AM hierarchy mirrored the linguistic phonological hierarchy of stressed syllables (~delta), syllables (theta), and onset-rime (beta/low gamma) units in the spoken material [8]. The number and the edge of bands are determined by principal component analysis (PCA) dimensionality reduction of original high-dimensional spectral and temporal envelope representations (Fig 1). This modulation filterbank can generate a cascade of amplitude modulators at different oscillatory rates, producing the AM hierarchy. On the other hand, the filterbank may partially introduce the artificial modulations into the stimuli because the bandpass filters can introduce modulations near the center-frequency of the filter through "ringing". Therefore, we also used a second AM-hierarchy extraction method (i.e., PAD) as a control. The PAD model does not implement the Hilbert transform, filtering, and PCA, but infers the modulators and a carrier based on Bayesian inference, identifying the envelope which best matches the data and the a priori assumptions. The PAD can be run recursively using different demodulation parameters each time, generate a cascade of amplitude modulators at different oscillatory rates (i.e., delta, theta, alpha, beta, and gamma), forming an AM hierarchy [50] (Fig 2).

In acoustic engineering, the music and speech signals could be expressed as the product of the slow-varying amplitude modulation (AM) patterns and the quickly-varying career or frequency modulation (FM) patterns [50,66]. That is, the AM patterns are associated with fluctuations in loudness or sound intensity, underlying a primary acoustic correlate of perceived rhythm based on onset timing, beat, accent, and grouping. In contrast, the FM patterns can be interpreted as

fluctuations in pitch and noise [50]. It is possible to separate the AM envelope of music from the FM structure by demodulation processes, the approaches adopted in the present study. We estimate the potential efficacy of signal-based mapping between acoustic structure and musical structure at each tier in the hierarchical modulation architecture.

4.1. Spectral Amplitude Modulation Phase Hierarchy (S-AMPH) model

(a) Signal processing: spectral and temporal modulations

The methodologies were based on a previous study by Leong and Goswami [8]. For the spectral modulation, the raw acoustic signal were passed through a 28 log-spaced equivalent rectangular bandwidth (ERB_N) filterbank spanning 100–7250 Hz, which simulates the frequency decomposition by the cochlea function in a normal human [23,49]. For further technical details of the filterbank design, see Stone and Moore [67]. The parameters of the ERB_N filterbanks and the frequency response characteristics are provided in S2 Appendix. Then, the Hilbert envelope was obtained for each of the 28 filtered-signals. Using the 28 Hilbert envelopes, the core modulation architecture was defined by principal component analysis (PCA). This can identify the appropriate number and spacing of non-redundant spectral bands, by detecting co-modulation in the high-dimensional ERB_N representation. For the temporal modulation, the raw acoustic signal was filtered into the number of spectral bands that were identified in the spectral PCA analysis. Then, the Hilbert envelope was extracted from each of the spectral bands. Further, the Hilbert envelope of each of the spectral bands were passed through a 24 log-spaced ERB_N filterbank spanning 0.9–40 Hz. Using the 24 Hilbert envelopes in each of the spectral bands, the core modulation architecture was defined by PCA. This can clarify co-activation patterns across modulation rate channels.

(b) PCA to finding the core modulation architecture in the high-dimensional ERB_N representation

In spectral PCA, the 28 spectral channels were taken as separate variables, yielding a total of 28 principal components. Only the top 5 principal components (PC) were considered for the further analysis, because these already cumulatively accounted for over 58% (on average) of the total variance in the original sound signal. In temporal PCA, the 24 channels in each of the spectral bands were entered as separate variables. Only the top 3 were considered for further analysis, because these cumulatively accounted for over 55% of the total variance in the original sound signal. Each of PC loading value was averaged across all samples in each genres (Western-classical music, Jazz, Rock, and

Children's song) and musical instruments (single-voice: Violin, Viola, Cello, and Bass; multi-voice: Piano and Guitar). The absolute value of the PC loading were used to avoid mutual cancellation by averaging an opposite valence across samples [8]. Then, peaks in the grand average PC loading patterns were taken to identify the core modulation architecture. Troughs were also identified because they reflect boundaries of edges between co-modulated clusters of channels. To ensure that there would be an adequate spacing between the resulting inferred modulation bands, a minimum peak-to-peak distance of 2 and 5 channels was set for the spectral and temporal PCAs, respectively. After detecting all the peaks and troughs, the core spectral and temporal modulation bands were determined based on the criteria that at least 2 of the 5 PCs and 1 of the 3 PCs showed a peak for spectral and temporal bands, respectively. On the other hand, the boundary edges between modulation bands were determined based on the most consistent locations of "*flanking*" troughs for each group of PC peaks that indicated the presence of a band. The detailed methodologies and examples are shown in Leong and Goswami [8].

4.2. Probability Amplitude Demodulation (PAD) model based on Bayesian inference

Amplitude demodulation is the process by which a signal (y_t) is decomposed into a slowly-varying modulator (m_t) and quickly-varying carrier (c_t):

$$y_t = m_t * c_t \qquad \ldots\ldots\ldots\

of the modulator and carrier is, given the observed signal. Thus, PAD summarizes the posterior distribution by returning the specific envelope and carrier that have the highest posterior probability and therefore represent the best match to the data. The PAD can be run recursively using different demodulation parameters each time, generate a cascade of amplitude modulators at different oscillatory rates (i.e., delta, theta, alpha, beta, and gamma), forming an AM hierarchy [50] (Fig 2). The positive slow envelope is modeled by applying an exponential nonlinear function to a stationary Gaussian process. This produces a positive-valued envelope whose mean is constant over time. The degree of correlation between points in the envelope can be constrained by the timescale parameters of variation of the modulator (envelope), which may either be entered manually or learned from the data. In the present study, we manually entered the PAD parameters to produce the modulators at each of five tiers of oscillatory band (i.e., delta: -4 Hz, theta: 4-8 Hz, alpha: 8-12 Hz, beta: 12-30 Hz, and gamma: 30-50 Hz). The carrier is interpreted as components including noise and pitches whose frequencies are much higher than the core modulation bands in phrase, prosodic, syllabic, phonological components. In each of music sample, the modulators (envelopes) of the five oscillatory bands were converted into the frequency domains by the Fast Fourier Transform (FFT). Spectral analysis of the modulator reflects how fast sound intensity fluctuates overtime. High modulation frequency corresponds to fast modulations and vice versa (see Fig 2). The modulation spectra were averaged across all samples in each genres (Western-classical music, Jazz, Rock, and Children's song) and musical instruments (single-voice: Violin, Viola, Cello, and Bass; multi-voice: Piano and Guitar).

4.3. Mutual information between different modulation bands

We also examined how a tier of the temporal hierarchy of music is interacted with another tier by mutual information analyses. Mutual information is a measure of the mutual dependence between the two variables. The mutual information can also be expressed as

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&= \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)}\right) - \sum_{x,y} p(x,y)\log p(y) \\
&= \sum_{x,y} p(x)p(y|x)\log p(y|x) - \sum_{x,y}\log p(y)\,p(x,y) \\
&= \sum_{x} p(x)\left(\sum_{y} p(y|x)\log p(y|x)\right) - \sum_{y}\log p(y)\left(\sum_{x} p(x,y)\right) \\
&= -\sum_{x} p(x)H(Y|X=x) - \sum_{y} p(y)\log p(y) \\
&= -H(Y|X) + H(Y) \\
&= H(Y) - H(Y|X)\ (bit) \qquad \text{..... (5)}
\end{aligned}
$$

where p(x,y) is the joint probability function of X and Y, p(x) and p(y) are the marginal probability distribution functions of the X and Y respectively, H(X) and H(Y) are the marginal entropies, H(X|Y) and H(Y|X) are the conditional entropies, and H(X,Y) is the joint entropy of X and Y [68].

This analysis could answer a question on how a certain oscillatory rhythm X (i.e., delta, theta, alpha and beta) is dependent on another oscillatory rhythm Y. Previous evidence suggests that the adjacent-band oscillators are not independent of, but interdependent on each others [46,47]. Given the evidence, we hypothesize that the adjacent tiers that can connect via so-called "*branch*" in the AM hierarchy are mutually dependent on each other compared with non-adjacent tiers. If so, the results may support a hierarchical "*tree-based*" structure of musical rhythm, highlighting the universality of AM hierarchy in music as well as speech.

To reveal this, we adopted the phase angle "θ" of the core temporal modulation envelopes corresponding to delta, theta, alpha and beta/gamma waves, which was detected in each of S-AMPH and PAD approaches. That is, in S-AMPH model, the 5 spectral envelopes (see Fig 1b) was passed through a second series of band-pass filters to isolate the 4 different AM bands based on the results of temporal PCA (channel edge frequencies: 0.9, 2.5, 7, 17 and 30 Hz). The phase angles were then calculated using each of the 4x5 temporal modulation envelopes. In PAD model, the phase angles were calculated using the four core modulators (envelopes) that have been detected in the last analyses, which correspond to delta, theta, alpha, and beta/gamma bands, respectively (for example, k, I, g, e in

Fig 2). Then, using the phase angle values, the mutual information between different temporal modulation bands was measured.

4.4. Phase Synchronization Analyses

Based on the findings of mutual information, we further investigated the integer ratios between "*adjacent*" AM hierarchies (i.e., the number of parent vs. daughter elements in an AM hierarchy). This analysis could answer a question on how many daughter elements a parent element encompasses in general. We adopted the core temporal modulation envelopes corresponding to delta, theta, alpha and beta/gamma waves, which was detected in each of S-AMPH and PAD approaches. That is, in S-AMPH model, the five spectral envelopes (see Fig 1b) was passed through a second series of band-pass filters to isolate the four different AM bands based on the results of temporal PCA (channel edge frequencies: 0.9, 2.5, 7, 17 and 30 Hz). In the end, the total numbers were 4x5 temporal modulation envelopes in the S-AMPH model. In contrast, in PAD model, we made use of the four core modulators (envelopes) corresponding to delta, theta, alpha, and beta/gamma bands, respectively (for example, k, I, g, e in Fig 2).

The Phase Synchronization Index (PSI) was computed between adjacent AM bands in the S-AMPH representation for each of the five spectral bands and in the PAD representation (i.e., delta vs. theta, theta vs. alpha, alpha vs. beta, beta vs. gamma phase synchronizations). The n:m PSI was originally conceptualized to quantify phase synchronization between two oscillators of different frequencies (e.g., muscle activity; [53]), and was subsequently adapted for neural analyses of oscillatory phase-locking [54]. For example, if the integer ratio is 1:2, then the parent element encompasses 2 daughter elements for the rhythm. The PSI was computed as:

$$PSI = |e^{1(n\theta 1 - m\theta 2)}| \qquad \ldots\ldots\ldots\

is perceived as being perfectly rhythmically regular (a repeating pattern of strong and weak beats), whereas a sound with a PSI of 0 is perceived as being random in rhythm.

To see if the detected findings properly represent universal characteristics of natural musical rhythm, we further conducted simulation analyses. We generated synthesized sounds that consisted of four temporal modulation envelopes (i.e., modulator) and one spectral frequency (carrier). That is, 2 Hz, 4 Hz, 8 Hz and 16 Hz sine waves were summarized to synthesize one compound tone waveform. The compound tone waveform was, then, multiplied by a 200 Hz sine waves. The synthesized waveform was assumed as a sound that includes temporal information of delta, theta, alpha and gamma rhythms, and spectral information of a pitch around to natural human voices (for the figure, see S5 Appendix). It is important to note that all of the temporal envelopes comprised of simple sine waves with frequencies of a power of 2. Hence, we can hypothesize that 1:2 integer ratios should clearly and consistently be appeared compared with other integer ratios. If the PSIs of natural music show different findings from those of artificial sounds, then the results may indicate that natural musical rhythm has covert universal integer ratios in an AM hierarchy.

Author Contributions

U.G. and T.D. conceived and conceptualized the study. T.D. developed the computational paradigm. T.D. collected and analyzed the data, prepared the figures, and wrote the original draft of the manuscript. U.G. reviewed and edited the manuscript. U.G. and T.D. finalized the manuscript.

Acknowledgments

This study was supported by Nakatani Foundation. The sponsor played no role in the study design nor in the collection, analysis, interpretation and writing up of the data.

Declaration of Interests

Competing financial interests: The authors declare no competing financial interests.

References

1. S. Mehr, M. Krasnow, G. A. Bryant, E. Hagen, Origins of music in credible signaling (2020).
2. M. Liberman, A. Prince, On stress and linguistic rhythm. *Linguist. Inq.* **8**, 249–336 (1977).
3. J. Mehler, *et al.*, A precursor of language acquisition in young infants. *Cognition* **29**, 143–178 (1988).
4. A. D. Patel, A neurobiological strategy for exploring links between emotion recognition in music and speech. *Behav. Brain Sci.* **31**, 589–590 (2008).
5. E. Selkirk, Phonology and Syntax. Tlie Relation between Sound and Structure (1984).
6. B. Hayes, *Metrical stress theory: Principles and case studies* (University of Chicago Press, 1995).
7. V. Leong, M. A. Stone, R. E. Turner, U. Goswami, A role for amplitude modulation phase relationships in speech rhythm perception. *J. Acoust. Soc. Am.* **136**, 366–381 (2014).
8. V. Leong, U. Goswami, Acoustic-Emergent Phonology in the Amplitude Envelope of Child-Directed Speech. *PLoS One* **10**, e0144411 (2015).
9. V. Leong, M. Kalashnikova, D. Burnham, U. Goswami, The Temporal Modulation Structure of Infant-Directed Speech. *Open Mind* **1**, 78–90 (2017).
10. A. Tierney, N. Kraus, Music training for the development of reading skills. *Prog. Brain Res.* **207**, 209–241 (2013).
11. D. Schön, *et al.*, Similar cerebral networks in language, music and song perception. *Neuroimage* **51**, 450–461 (2010).
12. I. Peretz, D. Vuvan, M.-É. Lagrois, J. L. Armony, Neural overlap in processing music and speech. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140090 (2015).
13. N. Kraus, E. Skoe, A. Parbery-Clark, R. Ashley, Experience-induced malleability in neural encoding of pitch, timbre, and timing. *Ann. N. Y. Acad. Sci.* **1169**, 543–557 (2009).
14. S. E. Trehub, E. E. Hannon, Infant music perception: Domain-general or domain-specific mechanisms? *Cognition* **100**, 73–99 (2006).
15. J. Morton, S. Marcus, C. Frankish, Perceptual centers (P-centers). *Psychol. Rev.* **83**, 405 (1976).
16. C. E. Hoequist Jr, The perceptual center and rhythm categories. *Lang. Speech* **26**, 367–376 (1983).
17. S. Scott, P-centres in speech-an acoustic analysis. *PhD thesis, Univ. Coll. London* (1993).
18. J. W. Gordon, The perceptual attack time of musical tones. *J. Acoust. Soc. Am.* **82**, 88–105 (1987).
19. M. Huss, J. P. Verney, T. Fosker, N. Mead, U. Goswami, Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex* **47**, 674–689 (2011).
20. E. O. Selkirk, The Role of Prosodic Categories in English Word Stress. *Linguist. Inq.* **11**, 563–605 (1980).
21. F. Lerdahl, R. Jackendoff, R. S. Jackendoff, *A Generative Theory of Tonal Music* (MIT Press, 1983).
22. V. Leong, Prosodic rhythm in the speech amplitude envelope: Amplitude modulation phase hierarchies (AMPHs) and AMPH models. *Cambridge, UK PhD Thesis* (2012).
23. B. C. J. Moore, *An introduction to the psychology of hearing* (Brill, 2012).

24. T. Dau, B. Kollmeier, A. Kohlrausch, Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *J. Acoust. Soc. Am.* **102**, 2906–2919 (1997).
25. X. Yang, K. Wang, S. A. Shamma, Auditory representations of acoustic signals. *IEEE Trans. Inf. theory* **38**, 824–839 (1992).
26. P. X. Joris, C. E. Schreiner, A. Rees, Neural processing of amplitude-modulated sounds. *Physiol. Rev.* **84**, 541–577 (2004).
27. S. Shamma, On the role of space and time in auditory processing. *Trends Cogn. Sci.* **5**, 340–348 (2001).
28. H. Luo, D. Poeppel, Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron* **54**, 1001–1010 (2007).
29. E. Ahissar, *et al.*, Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc. Natl. Acad. Sci.* **98**, 13367 LP-13372 (2001).
30. K. B. Doelling, L. H. Arnal, O. Ghitza, D. Poeppel, Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* **85**, 761–768 (2014).
31. B. Barton, J. H. Venezia, K. Saberi, G. Hickok, A. A. Brewer, Orthogonal acoustic dimensions define auditory field maps in human cortex. *Proc. Natl. Acad. Sci.* **109**, 20738–20743 (2012).
32. G. M. Di Liberto, J. A. O'Sullivan, E. C. Lalor, Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* **25**, 2457–2465 (2015).
33. K. B. Doelling, D. Poeppel, Cortical entrainment to music and its modulation by expertise. *Proc. Natl. Acad. Sci.* **112**, E6233–E6242 (2015).
34. A. L. Giraud, D. Poeppel, Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517 (2012).
35. M. J. Henry, J. Obleser, Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc. Natl. Acad. Sci.* **109**, 20095–20100 (2012).
36. S. Norman-Haignere, N. G. Kanwisher, J. H. McDermott, Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* **88**, 1281–1296 (2015).
37. S. Nozaradan, I. Peretz, M. Missal, A. Mouraux, Tagging the neuronal entrainment to beat and meter. *J. Neurosci.* **31**, 10234–10240 (2011).
38. T. Overath, J. H. McDermott, J. M. Zarate, D. Poeppel, The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* **18**, 903 (2015).
39. R. Santoro, *et al.*, Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.* **10**, e1003412 (2014).
40. N. Ding, L. Melloni, H. Zhang, X. Tian, D. Poeppel, Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci.* **19**, 158–164 (2016).
41. L. S. Baltzell, R. Srinivasan, V. Richards, Hierarchical organization of melodic sequences is encoded by cortical entrainment. *Neuroimage* **200**, 490–500 (2019).
42. H. Park, R. A. A. Ince, P. G. Schyns, G. Thut, J. Gross, Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Curr. Biol.* **25**, 1649–1653 (2015).
43. L. Fontolan, B. Morillon, C. Liegeois-Chauvel, A. L. Giraud, The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat. Commun.* **5**, 1–10 (2014).

44. D. Poeppel, The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time.’ *Speech Commun.* **41**, 245–255 (2003).
45. J. Obleser, C. Kayser, Neural Entrainment and Attentional Selection in the Listening Brain. *Trends Cogn. Sci.* **23**, 913–926 (2019).
46. J. Gross, *et al.*, Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLOS Biol.* **11**, e1001752 (2013).
47. P. Lakatos, *et al.*, An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *J. Neurophysiol.* **94**, 1904–1911 (2005).
48. R. E. Turner, M. Sahani, Demodulation as Probabilistic Inference. *IEEE Trans. Audio. Speech. Lang. Processing* **19**, 2398–2411 (2011).
49. B. R. Glasberg, B. C. J. Moore, Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* **47**, 103–138 (1990).
50. R. Turner, Statistical models for natural sounds. *Dr. Diss. Coll. London.* (2010).
51. W. Klein, R. Plomp, L. C. W. Pols, Vowel Spectra, Vowel Spaces, and Vowel Identification. *J. Acoust. Soc. Am.* **48**, 999–1009 (1970).
52. L. C. W. Pols, H. R. C. Tromp, R. Plomp, Frequency analysis of Dutch vowels from 50 male speakers. *J. Acoust. Soc. Am.* **53**, 1093–1101 (1973).
53. P. Tass, *et al.*, Detection of $\mathit{n}:\mathit{m}$ Phase Locking from Noisy Data: Application to Magnetoencephalography. *Phys. Rev. Lett.* **81**, 3291–3294 (1998).
54. B. Schack, S. Weiss, Quantification of phase synchronization phenomena and their importance for verbal memory processes. *Biol. Cybern.* **92**, 275–287 (2005).
55. R. Plomp, W. J. M. Levelt, Tonal consonance and critical bandwidth. *J. Acoust. Soc. Am.* **38**, 548–560 (1965).
56. N. Ding, *et al.*, Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **81**, 181–187 (2017).
57. U. Goswami, Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Ann. N. Y. Acad. Sci.* **1453**, 67–78 (2019).
58. M. J. McPherson, *et al.*, Perceptual fusion of musical notes by native Amazonians suggests universal representations of musical intervals. *Nat. Commun.* **11**, 2786 (2020).
59. D. A. Schwartz, C. Q. Howe, D. Purves, The statistical structure of human speech sounds predicts musical universals. *J. Neurosci.* **23**, 7160–7168 (2003).
60. E. Ladányi, V. Persici, A. Fiveash, B. Tillmann, R. L. Gordon, Is atypical rhythm a risk factor for developmental speech and language disorders? *Wiley Interdiscip. Rev. Cogn. Sci.*, e1528 (2020).
61. A. Bhide, A. Power, U. Goswami, A Rhythmic Musical Intervention for Poor Readers: A Comparison of Efficacy With a Letter-Based Intervention. *Mind, Brain, Educ.* **7**, 113–123 (2013).
62. R. Cumming, A. Wilson, V. Leong, L. J. Colling, U. Goswami, Awareness of Rhythm Patterns in Speech and Music in Children with Specific Language Impairments. *Front. Hum. Neurosci.* **9**, 672 (2015).
63. J. Araújo, S. Flanagan, A. Castro-Caldas, U. Goswami, The temporal modulation structure of illiterate versus literate adult speech. *PLoS One* **13**, e0205224 (2018).
64. Z. Kodály, The Selected Writings of Zolta´n Koda´ly. London: Boosey and Hawkes, [Lily Hala´py and Fred Macnicol, Trans.] (1974).
65. E. Jacques-Dalcroze, Rhythm, Music and Education [Rubinstein H, Trans.]. London: The Dalcroze Society Inc. (1980).

66. T. M. Elliott, F. E. Theunissen, The Modulation Transfer Function for Speech Intelligibility. *PLOS Comput. Biol.* **5**, e1000302 (2009).
67. M. A. Stone, B. C. J. Moore, Tolerable hearing aid delays. III. Effects on speech production and perception of across-frequency variation in delay. *Ear Hear.* **24**, 175–183 (2003).
68. T. Daikoku, Entropy , Uncertainty , and the Depth of Implicit Knowledge on Musical Creativity : Computational Study of Improvisation in Melody and Rhythm. **12**, 1–11 (2018).

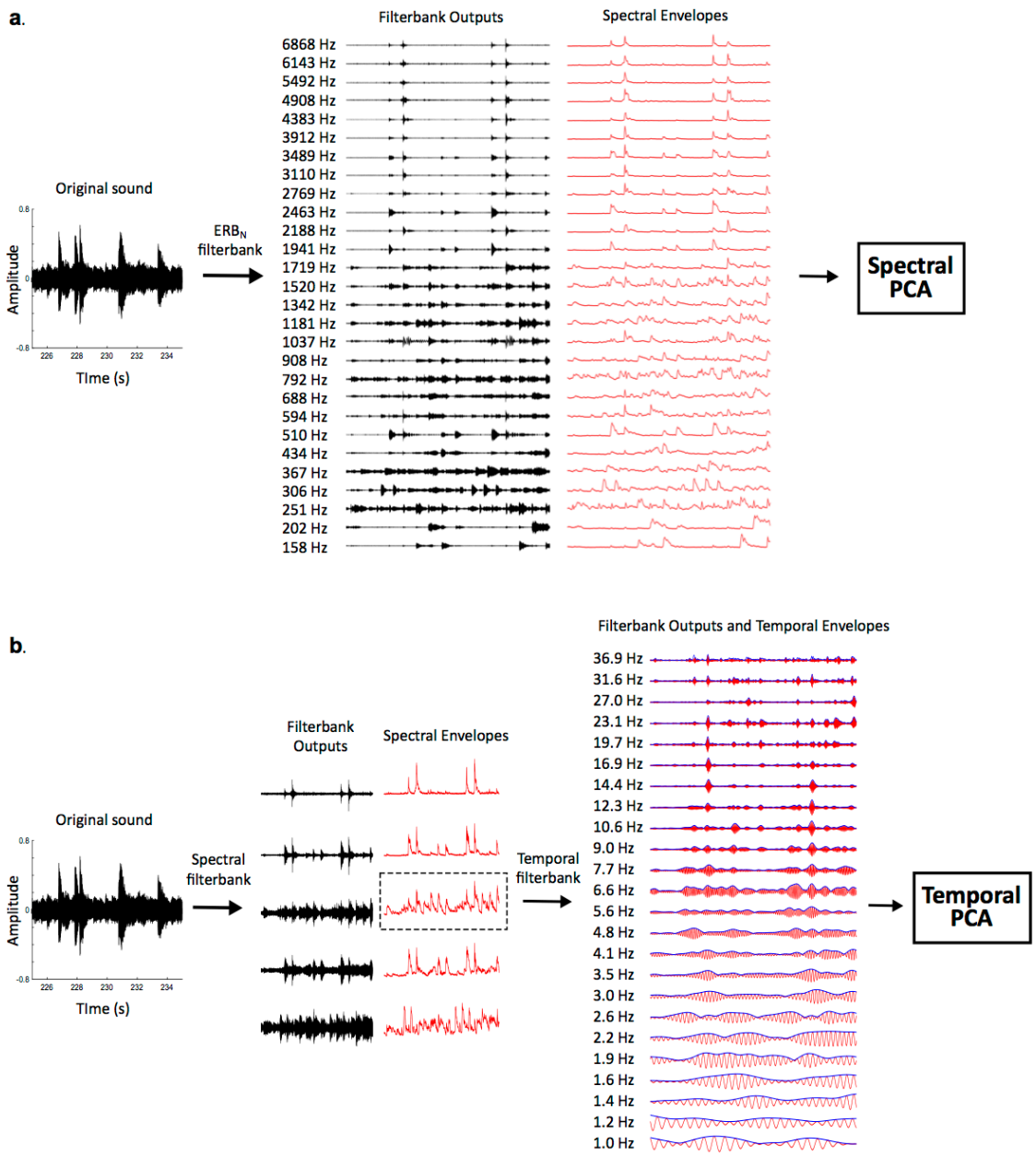

Fig 1. Signal processing steps in S-AMPH model. (**a**) The original sound is a part of a sample used in this study (Beethoven Piano Sonata No.14, Op27 No 2 "Moonlight"). Original sound is passed through an ERB_N-spaced filterbank, yielding a set of high-dimensional spectral channel outputs. The envelope is extracted from each spectral channel output using the Hilbert transform, and these envelopes are entered into the spectral PCA to identify patterns of covariation across spectral channels. (**b**) The original sound is passed through a low-dimensional spectral filterbank, yielding a small set of core spectral band outputs. The parameters of the low-dimensional spectral filterbank were determined in the Spectral PCA procedure (**a**). The envelopes are extracted from each spectral band output using the Hilbert transform. Each envelope is further passed through a high-dimensional modulation filterbank, yielding a set of high-dimensional modulation rate envelopes. This rate-filtering is performed for each spectral band envelope, but for simplicity, only the modulation rate envelopes from a single spectral band are shown in this figure. Finally, the power profiles of the modulation rate envelopes (bold blue line) are entered into a temporal PCA to identify patterns of covariation across modulation rates.

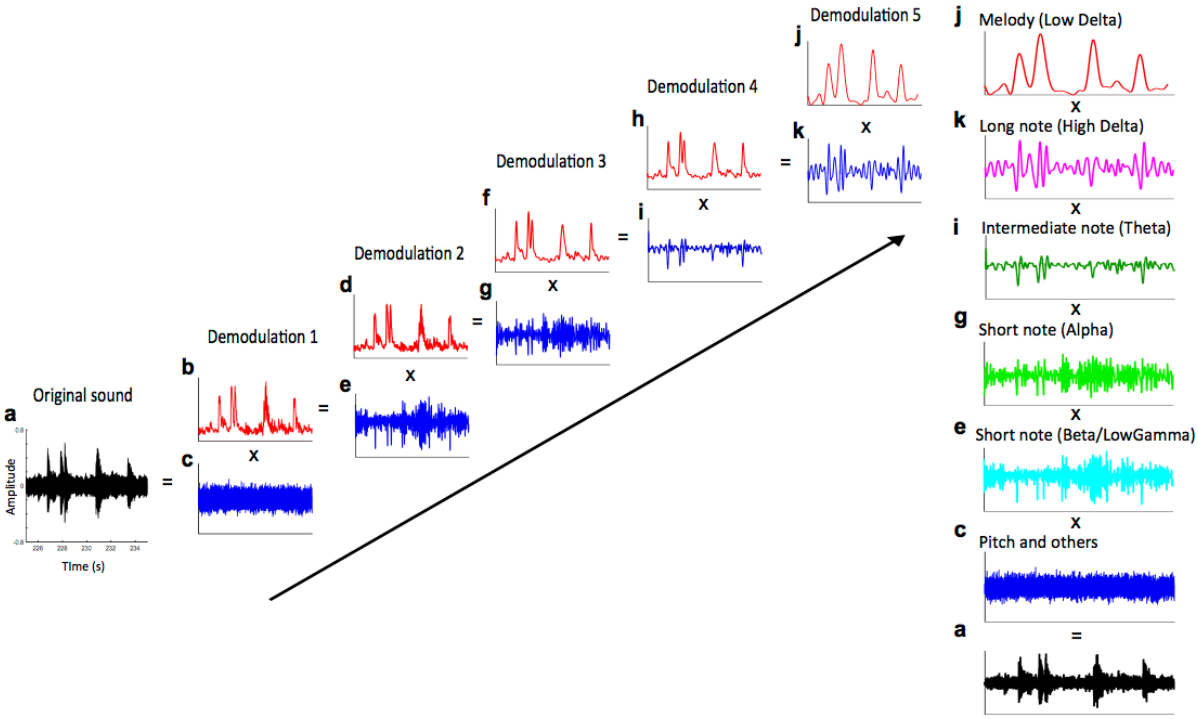

Fig 2. Signal processing steps in PAD model. Example of an AM hierarchy derived by recursive application of PAD. In the first demodulation round (left column), the data (**a**) are demodulated using PAD set to a fast time scale. This yields a relatively quickly-varying envelope (**b**) and a carrier (**c**). In the second demodulation round (middle column), the demodulation process is re-applied to the extracted envelope b, using a slower time scale than before. This yields a slower daughter envelope (**d**) and a faster daughter envelope (**e**). Daughter envelopes d and e form the two tiers of the resulting AM hierarchy (right column). Mathematically, these two tiers (**d** and **e**) can be multiplied back with the very first carrier (c, bottom left) to yield the original signal (**a**).

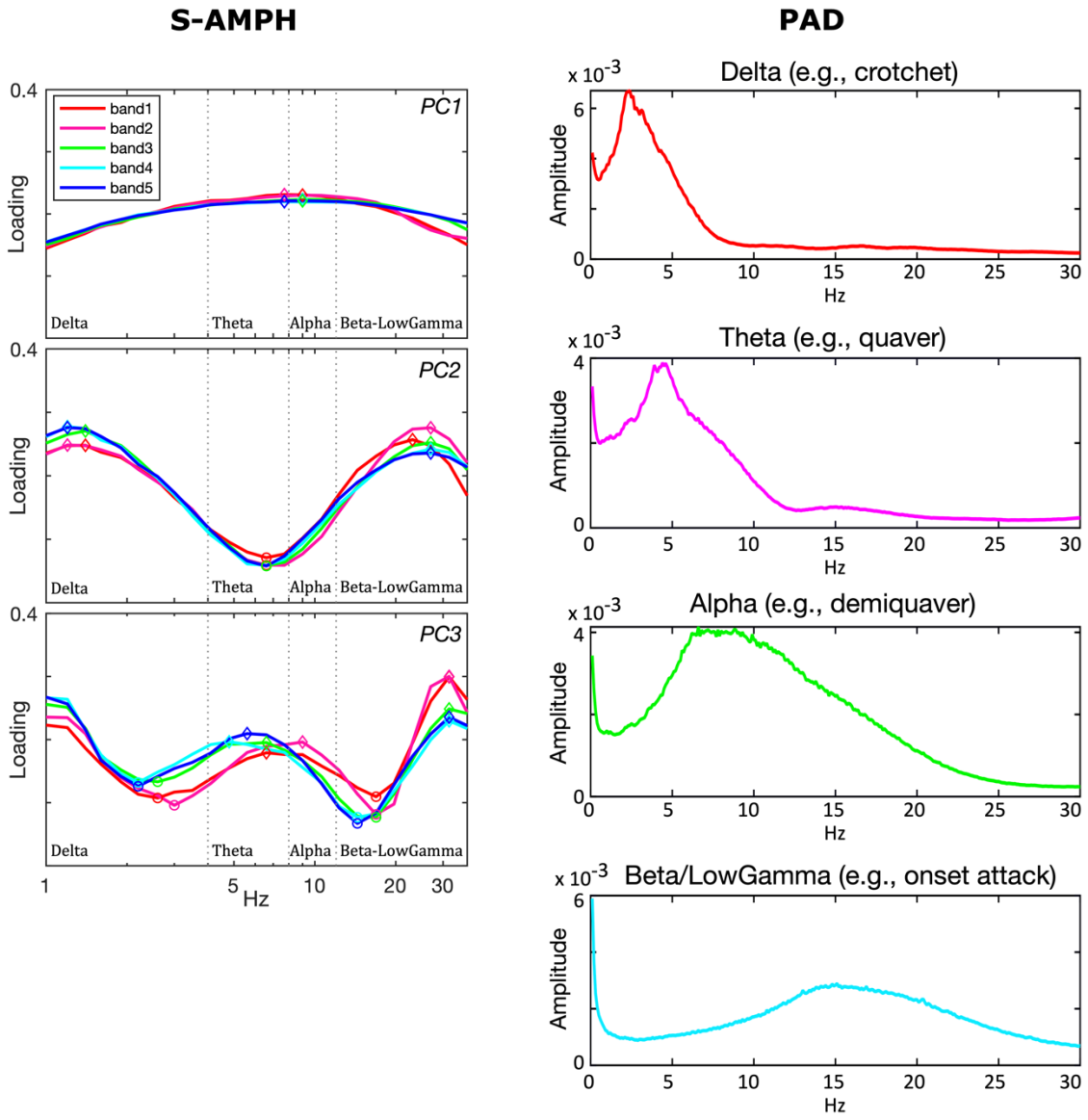

Fig 3. Core temporal modulation rates. Grand average temporal PCA component loading patterns in the S-AMPH (a) model and modulation spectra of FFT in the PAD model (b). Both models showed an AM hierarchy that consisted of delta-, theta-, alpha- and beta-rate AM bands.

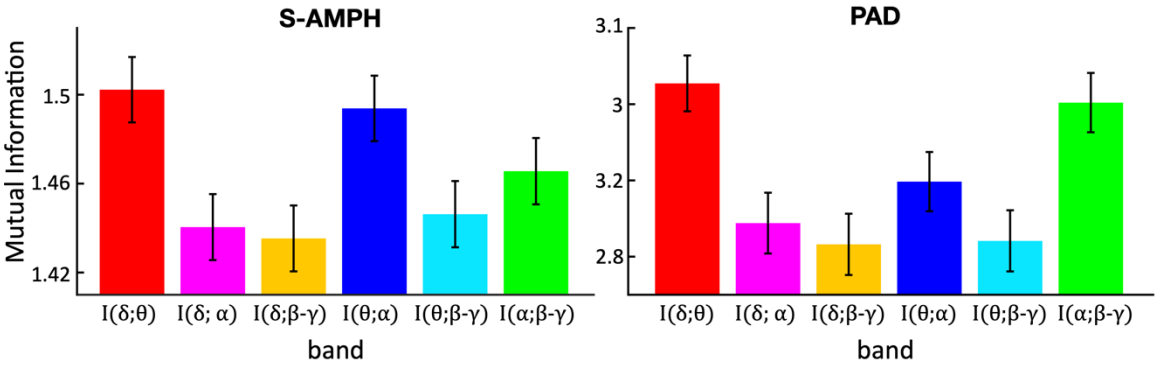

Fig 4. Mutual information between different tiers in an AM hierarchy. Both S-AMPH (a) and PAD models (b) showed stronger dependency between each adjacent tier of an AM hierarchy. Further, mutual dependence between delta- and theta-rate AM bands was the strongest of all mutual dependence in both models.

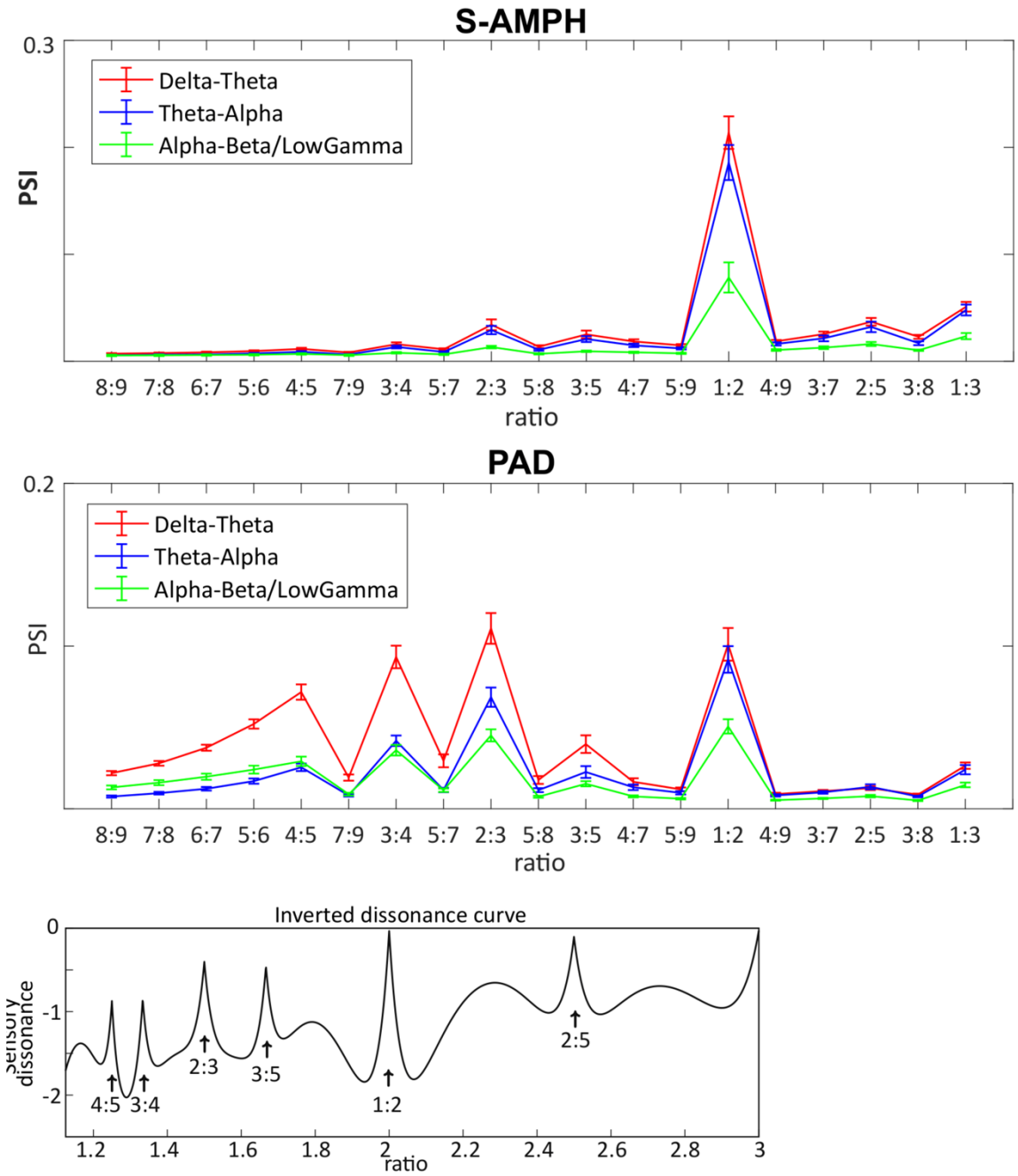

Fig 5. Phase synchronization index between different tiers in an AM hierarchy. Both S-AMPH (a) and PAD model (b) showed that the simpler integer ratios (i.e., m/n) synchronize their phase with each other. The inverted dissonance curve **(c)** was obtained by including the first five upper partials of tones with a 440 Hz (i.e., pitch standard, A4) fundamental frequency in calculating the total dissonance of intervals (55). It is of note that the peaks of PSI (a) correspond to those of the dissonance curve.

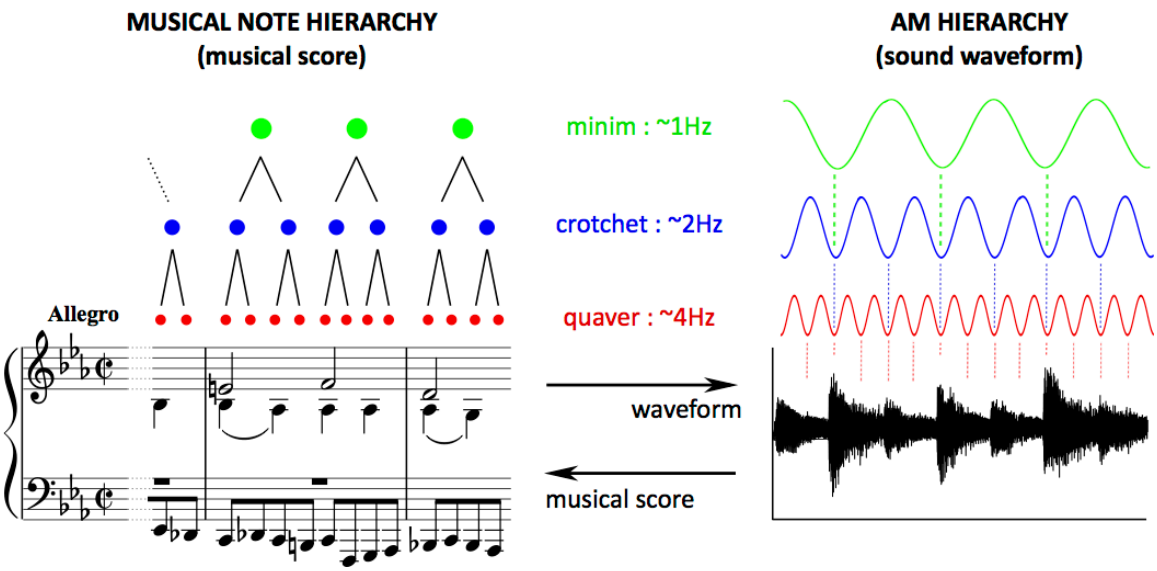

Fig 6. Hierarchical structure of rhythm in music. Left and right are the representation by musical score and the corresponding sound waveform of a part of the 33 Variations on a waltz by Anton Diabelli, Op. 120 (commonly known as the Diabelli Variations) by Ludwig van Beethoven. The musical rhythm could be hierarchically organised based either on amplitude modulations (AM, right) or on note values (left). As shown by the black lines in the musical note hierarchy (left) and the dotted vertical lines in the AM hierarchy (right), the adjacent tiers of the hierarchy (i.e., green & blue and blue & red) are mutually dependent on each other compared with non-adjacent hierarchical relations (i.e., green-red).

Table 1. Summary of the 4 temporal bands indentified from PCA by both S-AMPH and PAD.

Temporal bands	S-AMPH	PAD	Neural Oscillatory Rates
Band 1	0.9-2.5 Hz	-4 Hz	Delta
Band 2	2.5-7 Hz	4-8 Hz	Theta
Band 3	7-17 Hz	8-12 Hz	Alpha
Band 4	17-40 Hz	12-30 Hz	Beta/LowGamma

Supplemental Information Titles and Legends

S1 Appendix. Music Corpora (DOCX).

S2 Parameters of the ERB_N and modulation filterbanks, and the frequency response characteristics.

S3 Appendix. Individual Variation of PCA Loadings in S-AMPH model and those of FFT in the PAD model.

S4 Appendix. Individual Variation of Mutual Information.

S5 Appendix. Individual Variation of PSI in Each Integer Ratio.