Musical Rhythm across Western Genres arises from Hierarchical Temporal Modulation Structures that match the Structures found in Infant- and Child-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

Short Title: Temporal Modulation Structure of Music

* Corresponding author

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

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.

Abstract

Music and speech have structured rhythms that are related acoustically to slow temporal modulations in amplitude < ~40 Hz. The amplitude modulation (AM) timescales underlying music and speech have been thought to be distinct, with music showing a modulation peak at 2 Hz and music at 5 Hz. These differences in intrinsic structure have been argued to facilitate perceptual processing, however the comparisons rely on adult-directed speech (ADS). Analyses of infant- and child-directed speech (IDS, CDS) using amplitude envelope demodulation techniques yield modulation timescales that appear more similar to music, with a 2 Hz peak. Here we apply AM-driven phase hierarchy modelling approaches previously utilized with IDS and CDS to Western musical genres. We test whether the physical stimulus characteristics that yield speech rhythm in IDS and CDS may also describe the hierarchical rhythmic relationships that characterize music. Two models were applied. One utilized a low-dimensional representation of the auditory signal based on the human cochlear filterbank, and the second utilized probabilistic amplitude demodulation, estimating the modulator (envelope) and carriers using Bayesian inference. Both models revealed a very similar hierarchically-nested temporal modulation structure across musical genres and instruments, with core bands of AM and spectral patterning that matched prior analyses of IDS and CDS. Music showed similarly strong phase dependence between slower bands of AMs as IDS and CDS, critical to the perception of rhythm. Control analyses modelling other natural

sounds like wind and rain did not show similar temporal modulation structures and phase dependencies. We conclude that the AM timescales and their mutual dependencies in music and speech are not distinct, but highly similar. Acoustic rhythm in the musical speech directed to infants and in the musical genres invented by humans depends on perceiving hierarchical AM structures with systematically-coupled phase relations.

Author summary

The potential parallels between language and music have long fascinated researchers in social and cognitive science. Although language acquisition was once thought to require specialized neural architecture, it is now recognized that basic acoustic processing mechanisms are sufficient for infant language-learning, which begins with speech rhythm. Here we draw on our prior modelling work regarding the physical stimulus characteristics that yield speech rhythm in infant-directed speech (IDS), which has been described as sing-song speech, and rhythmic child-directed speech (CDS, English nursery rhymes) to propose a novel acoustic basis for musical rhythm. Adopting for the first time modelling approaches that enable decomposition of the 'amplitude envelope' of music and song, we reveal a systematic hierarchy of amplitude modulations and their phase relations underpinning musical rhythm. We demonstrate that the temporal modulation hierarchy in music shares fundamental commonalities with IDS and CDS, but not with other natural sounds (e.g., wind). As also found in IDS and CDS, mutual phase dependence between delta- and theta-rate amplitude modulation bands appears core to metrical structure. The striking match between the temporal modulation structure of music and IDS suggests that music is likely encoded by the brain using the same neural oscillatory mechanisms that encode speech.

Introduction

Hierarchical Temporal Modulation Structure of Auditory Rhythm

Music and language are ubiquitous in human societies [1]. Rhythmic structure is a fundamental feature of both domains, and involves sequences of events (such as syllables, notes, or drum beats) which have systematic patterns of timing, accent, and grouping [2]. Infant-directed speech (IDS, also called baby talk or Motherese) has been described as sing-song speech, and has particular prosodic characteristics that have been suggested to explain both natural selection for human language from an anthropological perspective [3], and to facilitate infant learning of the phonological structure of human languages [4]. Recently, a novel acoustic perspective on child language acquisition, Temporal Sampling theory [5,6] has led to the modelling of both IDS and rhythmic child-directed speech (CDS) such as nursery rhymes in terms of patterns of amplitude modulation (AM) in the amplitude envelope [7,8,9]. The amplitude envelope of any sound is the slower changes in AM (intensity or signal energy) that unfold over time. It is possible to separate the AM characteristics of speech from the frequency modulation (FM) characteristics by acoustic engineering methods for decomposing the amplitude envelope (called demodulation, [10]). Recent demodulation modelling of IDS and CDS revealed (a) that the modulation peak in IDS is ~2 Hz, in direct contrast to the modulation peak found for adult-directed speech (ADS, ~5 Hz) [7], (b) that rhythmic patterning is represented in the phase

relations of AM bands in the amplitude envelope, particularly those AM bands corresponding temporally to delta, theta and beta/low gamma bands in electroencephalography (EEG) [9], and (c) that linguistic rhythm patterns perceived by adult listeners, such as trochaic and iambic meters, depend on the temporal alignment of modulation peaks in the delta-rate and theta-rate bands of AM in the amplitude envelope [8]. This 'phase hierarchy' modelling approach (S-AMPH, Spectral-Amplitude Modulation Phase Hierarchy, see [9]) also revealed important differences in the mutual dependencies between AM bands in the temporal modulation structure of IDS versus ADS. For example, a direct comparison of IDS and ADS using the same speakers reported by [7] showed that the phase synchronisation (temporal alignment) of delta- and theta-rate AM bands was significantly greater in IDS than in ADS, while the phase synchronisation of theta- and beta/low gamma rate AM bands in ADS was significantly greater than in IDS. The low-frequency phase synchronisation of AM bands in ADS is also significantly greater for spontaneous speech spoken by literate adults than for spontaneous speech spoken by illiterate adults [11]. Accordingly, the temporal modulation structure of speech varies depending on whether an infant or an adult is being addressed, as well as with the literacy level of the speaker. This raises the question of which kind of human speech should be used as a basis for extracting the physical stimulus characteristics that are to be compared with music. It is notable that prior music/speech analyses of the temporal modulation structure of speech, across languages, have been based on highly literate adult speakers, typically university students [12].

It is plausible that the physical stimulus characteristics that describe IDS may also describe the hierarchical rhythmic relationships that characterize modern music. The strong phase synchronisation of delta- and theta-rate AMs in IDS reflects stronger rhythmic synchronization and acoustic temporal regularity, contributing to the 'sing-song' nature of Motherese remarked by [3]. Regarding the physical signal, it is likely that the stronger phase relations that characterize IDS represent structural acoustic differences from ADS that are important for early language learning. The strong AM phase relations of IDS also underpin its prosodic structure, which is thought to have played a key role in the pre-linguistic foundations of the protolanguage(s) evolved by early hominins [3]. This prosodic structure appears to have arisen from hominid soothing routines, devised to settle infants placed on the ground during foraging behaviour: 'affectively positive, rhythmic melodies' [3]. Intriguingly, current analyses of the lullabies sung by mothers to their infants across cultures reveals a beat rate of ~2 Hz (120 bpm, see [13]), matching the modulation peak found in studies of IDS [7] as well as music [12]. Human song, and possibly also human music across genres, may thus exhibit similar physical stimulus characteristics to IDS.

Music shares many commonalities with speech at both an acoustic and cognitive level [14-17]. The human peripheral auditory system responds to 'broadband' sounds like speech or music by filtering the sounds into narrowband signals with a wide range of center frequencies [18]. This filtering occurs in the human cochlea, and each of the resultant narrowband signals can be modelled as a rapidly oscillating carrier signal with a relatively slowly varying amplitude

envelope (the cochlear filterbank model of human hearing [19, 20]). Demodulation approaches to characterizing the physical stimulus structure of IDS and CDS decompose the amplitude envelope into these narrowband signals, and then seek systematic patterns of AM [7,9]. Like pre-verbal infants listening to spoken language, the music listener experiences a coherent perceptual signal with accentuation and rhythms that are captured by the overall amplitude envelope. The music listener needs to identify discrete units to gain meaning, like musical notes and phrasing, analogous to infants needing to identify discrete units like syllables, words and syntactic phrases from the prosodic rhythm structure of IDS. The systematic patterns of AM nested in the amplitude envelope of both IDS and CDS, in separable spectral bands, support the identification of these discrete units. For example, application of the S-AMPH to English nursery rhyme corpora showed that the model identified 72% of stressed syllables correctly, 82% of syllables correctly, and 78% of onset-rime units correctly if a particular AM cycle was assumed to match a particular speech unit [9]. If the nursery rhymes were chanted to a regular 2 Hz beat, then the model identified over 90% of each type of linguistic unit. Accordingly, decomposition of the amplitude envelope of different musical genres may identify similar hierarchical AM structures in predictable spectral (pitch) bandings that provide a perceptual basis for perceiving rhythm patterns, musical notes and musical phrasing. Whether music and IDS/CDS would exhibit similar salient bands of AMs, similar spectral banding and similar phase dependencies between AM bands is currently unknown.

Here we explore these questions using the S-AMPH model, a set of signal processing algorithms initially developed for English nursery rhymes, one common form of CDS [9,21]. The S-AMPH is a set of demodulation algorithms which take into account the cochlear filterbank model of human hearing. The S-AMPH then computes the statistical structure of the acoustic signal by decomposing the amplitude envelope. To date, the S-AMPH has been applied to IDS, CDS and ADS [7-9, 11]. As well as applying the S-AMPH to music, we also tested our hypothesis using Probabilistic Amplitude Demodulation or PAD [22]. PAD is a brain-neutral approach that models the statistical structure of acoustic inputs based purely on Bayesian inference of likely envelopes and carriers (spectral bands). In our prior work with English nursery rhymes [8], both S-AMPH and PAD modelling showed that adult perception of linguistic rhythm patterns such as trochaic and iambic meters depended on the temporal alignment of modulation peaks in the delta- and theta-rate bands of AM in the amplitude envelope. Accordingly, we predict that musical meter may also depend primarily on the temporal alignment of modulation peaks in the delta- and theta-rate bands of AM, across musical genres. To test the specificity of this prediction regarding speech and music, we also modelled other quasi-rhythmic natural sounds such as wind and rain, using both the S-AMPH and PAD approaches.

The theoretical framework underpinning our novel modelling approach is Temporal Sampling theory (TS theory [5]). TS theory was initially developed to provide a systematic sensory/neural/cognitive framework for explaining childhood language disorders, with a view to supporting musical remediation of such

disorders [23]. TS theory proposes that accurate sensory/neural processing of the amplitude envelope of the speech signal is one foundation of language acquisition, particularly regarding the AM rates <10 Hz that carry prosodic and rhythmic information [24]. A range of studies across languages support the view that sensory/neural processing of these slower AM rates is impaired in children with developmental disorders of language [6]. Children with developmental disorders of language also show impaired discrimination of amplitude rise times. Amplitude rise times, the rate of change of AMs at different frequencies, are important for the perception of rhythmic timing in both music and speech, as they determine the acoustic experience of ‘P-centres’. P-centres are the perceptual moment of occurrence (‘perceptual centre’) of each musical beat or syllable for the listener [25,26], and amplitude rise times are typically called attack times in the musical literature [27,28]). Perceiving the beat structure in both music and speech is known to be developmentally inter-related [29]. Children’s prosodic perception and their musical rhythm perception are both related to individual differences in amplitude rise time perception, and children's performance in a musical beat perception task uniquely predicts their performance on tests of phonological awareness, both concurrently and longitudinally [29,30]. Amplitude rise times also provide sensory landmarks that *automatically* trigger brain rhythms and speech rhythms into temporal alignment, acting as acoustic landmarks that phase-reset ongoing neural oscillatory activity [31]. According to TS theory, this neural phase-resetting process is atypical in developmental

disorders of language, and may be improved by interventions that link musical and speech rhythms for children [23].

Our prediction that the perception of musical meter may depend on the temporal alignment of different AM bands across musical genres also relates to linguistic theory [32-34]. Linguistically, hierarchical structures like the phonological hierarchy of prosodic, syllabic, rhyme and phoneme levels nested within speech rhythm are classically represented as a tree that captures the *relative prominence* of units [14,34]. Such tree representations may also provide a good model regarding the core principles of metrical structure in music [35]. 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. To give an example from CDS, a parent node such as the trisyllabic word “pussycat” in the nursery rhyme “Pussycat pussycat where have you been”, which is also the prosodic foot, would have 3 daughter nodes at the next hierarchical level, comprising the three syllables. S-AMPH modelling of nursery rhymes such as this revealed that the acoustic underpinning of this hierarchical metrical structure was an AM phase hierarchy that was consistent across the spectral range of speech (collapsed by principal component analysis or PCA into 5 bands) which comprised 3 temporal bands of AMs also derived using PCA (AMs in a ’delta’ band, 0.9 – 2.5 Hz centred on ~2 Hz; AMs in a ’theta’ band, 2.5 – 12 Hz centred on ~5 Hz; and AMs in a ’beta/low gamma band’, 12 – 40 Hz centred on ~20 Hz, [9]). English nursery rhymes with different metrical

structures like “Jack and Jill went up the hill” (trochaic rhythm), “As I was going to St Ives” (iambic rhythm) and “Pussycat pussycat where have you been” (dactyl rhythm) all showed the same acoustic hierarchical AM structure. Which metrical structure was perceived by the listener depended on the temporal alignment of AM peaks in the delta- and theta-rate AM bands [8]. For example, perceptual experience of a trochaic rhythm could be changed to an iambic rhythm by phase-shifting the AM peaks in the delta-rate AM band by 180’, so that they now aligned with different AM peaks in the theta-rate AM band compared to their alignment prior to the phase shift. It should be noted that the terms ‘delta-rate’ and ‘theta-rate’ AM bands were adopted to describe the results of the speech demodulation analyses because TS theory was based in part on the neural oscillatory bands that track human speech in adult cortex [36-42], described further below. The AM bands identified by the S-AMPH in both IDS and CDS equate temporally to electrophysiological rhythms found across the brain at the oscillatory rates of delta, theta and beta-low gamma [7,9]. To date, despite a number of approaches encompassing brain-based analyses [43-46], no studies have examined the temporal correlates of musical rhythm from this hierarchical AM modelling perspective. Note that our modelling approach is conceptually distinct from models that identify the tactus or beat markers in singing [47], models of pulse perception based on neural resonance [48], and oscillatory models of auditory attention based on dynamic attending [49] via its focus on decomposition of the amplitude envelope of the acoustic input.

Biological Underpinning: Neural Oscillators

Human speech perception relies in part on neural tracking of the temporal modulation patterns in speech at different timescales simultaneously and then binding them into a single percept, ‘multi-time resolution processing’ [36-38, 50]. This neural tracking (also described as aligning, phase-locking or entraining) relies on oscillatory cortical activity, and for both music and speech oscillatory activity is known to align with selected rhythmic features of the input such as syllables or musical beats [51-53]. Indeed, if musical rhythms and language rhythms are designed to be identical in a particular stimulus set (achieved by creating matching hierarchical rhythmic structures in words in German sentences and Waltz-like 3-count musical pieces), then EEG recordings show similar phase-locked neural responses in listening adults [54]. Accordingly, it is biologically plausible to suggest that at a more global ‘Gestalt’ perceptual level, rhythm perception in music and language may depend on neural entrainment to the AM hierarchies nested in the amplitude envelope of music versus speech respectively. Most of the slow energy modulations within the speech amplitude envelope reflect intensity patterns associated with syllable production [55]. However, within the overall speech envelope there are many amplitude envelopes of the different constituent (spectral) frequencies changing at different temporal rates. For speech, this ‘modulation structure’ is modelled by decomposing the amplitude envelope using approaches such as PAD and the S-

AMPH. The resulting temporal modulation spectrum has a relatively straightforward neurophysiological interpretation which could also apply to musical signals. As noted earlier, auditory psychophysics has established that the cochlea decomposes acoustic signals into narrow frequency bands [18]. These narrowband signals are then low-pass filtered at different stages of the auditory pathway [56], thereby extracting the temporal modulation envelope of the narrowband signals [57]. Oscillatory cortical networks track the temporal modulation envelope < ~40 Hz [58,59], with different oscillators (delta (<4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), gamma (30–80 Hz)) phase-synchronizing with different AM patterns at matching rates. Amplitude rise times are used as automatic triggers for this phase re-setting, so that amplitude envelopes at different frequencies in the signal become temporally aligned with neural oscillators at these frequencies [31,36,37]. These populations of auditory neurons thereby encode the modulation spectra of the temporal envelope at different neural signalling rates [36-41, 53, 60,61].

This neural ('speech-brain') alignment has been shown to contribute to parsing of the speech signal into phonological units such as syllables and words [41], and potentially to parsing musical units such as such as notes, beat, and meter [46]. For language, delta, theta, and beta/gamma oscillators in auditory cortex contribute to the perception of prosodic, syllabic, and phonetic information respectively [36,42,62]. Of core relevance to demodulation modelling approaches, there is also growing biological evidence that these adjacent-band oscillators are not independent of, but interdependent on, each other [52,63]. For

example, the phase of delta oscillators modulates the phase of theta oscillators, and theta phase modulates beta/gamma power [52]. Accordingly, there are mechanistic phase dependencies in the neural system, which mirror the acoustic phase dependencies between AM bands revealed by the S-AMPH for IDS and CDS. Thus, musical rhythm may depend on the same acoustic phase dependencies. It has already been shown that neural phase locking to periodic rhythms present in musical tempi is selectively enhanced compared to frequencies unrelated to the beat and meter [45]. However, to date the amplitude envelope of different musical inputs has not been decomposed in order to discover whether beat and meter are systematically related to adjacent bands of AMs that are physically connected by mutual phase dependencies. Accordingly, the aim of the present study is to explore the physical characteristics of music across Western musical genres and instruments utilising an AM phase hierarchy modelling approach to decompose the amplitude envelope of each genre. It is expected that music may induce phase-locking and synchronisation of neural oscillators to temporal modulation patterns that match the temporal modulation patterns found in CDS and IDS. By hypothesis, phase locking to different bands of AM present in the amplitude envelope of each genre may enable parsing of the signal to yield the perceptual experience of musical components such as minim, crotchet, and quaver (half, quarter, and eighth notes). The acoustic structure of the amplitude envelope should also contribute systematically to the perceptual experience of beat, tempo, and musical phrasing.

The music samples used to test these hypotheses 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 previously applied to CDS, the S-AMPH, [9], and PAD, [10,22]). Both models parse the amplitude envelope of the music signals into an hierarchy of AM bands, but the principles underpinning their operation are different. The S-AMPH simulates the frequency decomposition known to be carried out by the cochlea [18,64,65], thereby aiming to decompose the amplitude envelope of music in the same way as the human ear. PAD infers the modulators and carriers in the envelope based purely on Bayesian inference, thereby carrying out amplitude demodulation on a neutral statistical basis that makes no adjustments for the human hearing system (see Methods). We hypothesized that the two models would show consistent findings regarding hierarchical AM patterns across musical genres, as this was previously the case for CDS [8]. Accordingly, both modelling approaches (filtering vs probabilistic) should demonstrate the same core acoustic principles underlying the structure of musical rhythm. In particular, we hypothesized that the core temporal modulation hierarchy would be similar across Western music and infant-directed language. Given the biological evidence that each neural oscillator modulates the adjacent-band oscillator during speech perception [52,63], and our prior acoustic modelling data with IDS, CDS and ADS revealing significantly greater delta-theta

AM band phase alignment in IDS and CDS, we also hypothesized that the adjacent tiers in the temporal hierarchies of music would be highly dependent on each other compared with non-adjacent tiers, particularly for delta-theta AM coupling. To test whether these predicted commonalities across speech and music would be unique to these two culturally-acquired systems, we also used both the PAD and S-AMPH models to examine mutual dependency in natural sounds with quasi-rhythmic content such as rain, wind, fire, storms and rivers. These natural sounds have also been present since early hominid times, but their statistical structure has not been constrained by the human brain.

Signal Processing

The S-AMPH model was used to create a low-dimensional representation of the musical signals, using equivalent rectangular bandwidth (ERB_N) filterbanks [18,65]. This 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 structures in the amplitude envelope for each musical genre and instrument (see S2 Appendix). The number and the edge of bands were determined by PCA dimensionality reduction of original high-dimensional spectral and temporal envelope representations (Fig 1). This modulation filterbank generated a cascade of amplitude modulators at different oscillatory rates, producing the AM hierarchy.

The PAD model did not implement the Hilbert transform, filtering, and PCA, but inferred the modulators and a carrier based on Bayesian inference, identifying the envelope which best matched the data and the *a priori* assumptions. PAD was 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 [22] (Fig 2). PAD was applied to each musical genre and instrument.

Results

Amplitude Modulation Properties of Western 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 Methods). The PCA has previously been used for dimensionality reduction in speech studies (e.g. [66,67]). The present study focused on the absolute value of component loadings rather than the component scores. The loadings indicate the underlying patterns of correlation between high-dimensional channels. That is, 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 belonged 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. 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. The loading patterns for each Western music genre (Western classical symphonies, jazz, rock, children's songs) and instrument (piano, guitar, cello, bass, viola, violin, single voice, multiple voices) are depicted in S3 Appendix. Next, peaks and troughs in the grand average PC loading patterns were identified because they reflected centres 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 grand average as well as the loading patterns and cumulative contribution ratios for each music genre and instrument. 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 were 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). Further peaks and troughs were identified providing evidence for four further spectral bands, please see S3 Appendix and Methods. Based on the a priori criteria (see Methods), the spectral PCA thus 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, which were consistent across musical instruments and the human voice, are proportionately-scaled with respect to the logarithmic frequency sensitivity of human hearing. As predicted, these results are similar to the spectral bands previously revealed by modelling IDS and CDS [7,8,9]. It can also be noted that the loading patterns for the 5 PCA components showed roughly similar characteristics across the genres. There was some individual variation at each spectral modulation band, see Appendix S3.

Temporal PCA. S3 Appendix also shows the temporal loading patterns and cumulative contribution ratios for each music genre and each instrument, while 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 S3). 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 resulted 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 were 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 a priori criteria (Methods), the

temporal PCA thus provided evidence for the presence of 4 core bands with 3 boundaries across the different musical genres and instruments (see Table 1). Perceptually, cycles in these AM bands may yield the experience of crotchets, quavers, demiquavers and onsets, as shown in Table 1.

These AM bands in music matched those previously found in IDS, but the AM bands in the nature sounds did not (PC3 in Fig 4). In particular, the strong peaks in the delta and theta bands, along with the strong flanking trough between these bands, are clearly visible for music and speech compared with the nature sounds. As predicted, therefore, the results of the temporal PCA match prior studies of CDS and IDS [7,8,9].

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

Temporal bands	S-AMPH	PAD	Neural Oscillatory Rates	Note value
Band 1	0.9-2.5 Hz	-4 Hz	Delta	Crotchet (quarter note)
Band 2	2.5-7 Hz	4-8 Hz	Theta	Quaver (eighth note)
Band 3	7-17 Hz	8-12 Hz	Alpha	Demiquaver (sixteenth note)
Band 4	17-40 Hz	12-30 Hz	Beta/LowGamma	Onset attack

Amplitude Modulation Properties of Music from PAD

PAD ignores the logarithmic frequency sensitivity of human hearing, implementing amplitude demodulation by estimating the most appropriate modulator (envelope) and carrier based on Bayesian inference (for more detail, see Methods). Accordingly, PAD provides a good test of the assumption that there is a systematic hierarchy of temporal modulations underpinning both Western music and (English) IDS. Further, PAD is exempt from the possibility that the filterbank used in the S-AMPH may have partially introduced artificial modulations into the stimuli (because bandpass filters can introduce modulations near the center-frequency of the filter through “ringing”). 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 [22] (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 regardless of any tempo variations (such as speeding up or slowing down) that may depend on the performer or the particular music. Accordingly, it determines the frequencies that comprise the core temporal modulation structure of each musical genre. The carrier was interpreted as components including noise and pitches whose

frequencies were 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 Fast Fourier Transform (FFT). We had hypothesized that modelling with PAD should reveal the same core principles as the S-AMPH modelling regarding hierarchical AM patterns.

S3 Appendix shows the grand average for the modulation spectra of FFT as well as the loading patterns and cumulative contribution ratios for each music genre and instrument along with individual variation. Overall, low spectral frequencies (~5Hz) showed higher consistency across musical genres and instruments while mid-to-high spectral frequencies showed some individual variation, as also found for S-AMPH. Fig 3 represents the grand average modulation spectra of FFT for each oscillatory band. Four peak frequencies were detected at ~2.4 Hz, ~4.8 Hz, ~9 Hz and 16 Hz. Accordingly, PAD yielded similar findings to the S-AMPH model, detecting peak frequencies in AM bands corresponding temporally to neural delta, theta, alpha and beta/gamma neural oscillatory bands (see Table 1). As predicted, the modelling suggests the same core principles of AM structure across musical genres as the S-AMPH model. Accordingly, the AM structure of music shown in Table 1 is found irrespective of the modelling approach adopted.

Mutual Information in Both Models

To examine whether mutual dependencies between AM bands in the temporal modulation structure of different musical genres was more similar to the dependencies identified in IDS rather than in ADS, a mutual information analysis method was employed. As noted earlier, prior modelling of IDS and CDS has revealed a significantly higher phase dependency between delta- and theta-rate AM bands compared to ADS. ADS by contrast shows a significantly higher phase dependency between theta- and beta/low gamma rate AM bands compared to IDS. Mutual phase dependency between tiers of the AM bands in music was assessed using mutual information (MI) analyses. MI enables computation of whether a certain AM rhythm is dependent on another AM rhythm. Previous evidence suggests that adjacent-band neural oscillators are not independent of, but rather interdependent on, each other [52,63]. Accordingly, we had 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*" acoustic AM structure underlying musical rhythm, highlighting the systematic presence of the AM hierarchy in Western music as well as in speech directed to pre-verbal infants.

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 MI for each music genre and instrument, revealing high consistency between Western music genres and instruments in both S-AMPH and PAD models. Therefore, to probe potentially ubiquitous mutual dependencies across Western genres and instruments, we focused on the grand average (Fig 5). The results showed that adjacent tiers of the AM hierarchy were 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 matched the results of the prior speech-based modelling with IDS and CDS rather than ADS [7,8,9].

As an additional test, we also examined the MI between AM tiers of natural quasi-rhythmic non-musical inputs experienced by humans (utilising sounds such as rain, wind, fire, storms and rivers; see S1 Appendix), using both the S-AMPH and PAD models. The results were notably different to music. Inspection of the figures in S4 Appendix shows that compared with music, the mutual dependence between delta- and theta-rate AM bands of natural sounds

was similar to non-adjacent tiers (S4 Appendix). The current modelling suggests that for music, delta-theta phase alignment of AM bands underpins metrical structure, at least for Western musical genres. Accordingly, metrical structure, a feature shared by both music and speech, depends on the same core delta-theta AM phase relations in both domains.

Multi-Timescale Phase Synchronization in both Models

The demonstration of mutual dependency does not by itself capture metrical structure, as each AM cycle at a particular timescale may encompass one or more AM cycles at a faster timescale. To identify how many daughter elements a parent element could encompass in general, we next investigated the integer ratios between "*adjacent*" AM bands. For example, if the integer ratio is 1:2, then the parent element encompasses 2 daughter elements for the rhythm. An example from speech would be a tongue twister like "Peter Piper picked a peck of pickled peppers", which follows a 1:2 ratio (two syllables in each prosodic foot). To assess the integer ratios for each pair of mutually dependent AM bands in our selected musical genres, a phase synchronization index (PSI) was computed between adjacent modulation rate bands. The PSI was computed using both the S-AMPH and PAD representations, using the five spectral bands in S-AMPH and four in PAD (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; [68]), and was subsequently adapted for neural analyses of oscillatory phase-locking [69].

S5 Appendix shows the PSI for each music genre and instrument, revealing high consistency between music genres in both S-AMPH and PAD models. Further analysis focused on the grand average (shown in Fig 6). 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 were also higher than the other integer ratios, suggesting that the simpler integer ratios (i.e., m/n) were likely to synchronize in phase with each other. For spoken languages, the m/n ratio between two adjacent AM bands tends to vary with linguistic factors such as how many phonemes typically comprise a syllable (e.g. 2 phonemes per syllable for a language with a consonant-vowel syllable structure like Spanish, hence a theta-beta/low gamma PSI of 1:2, but 3 phonemes per syllable for a language with largely consonant-vowel-consonant syllable structures like English, hence a theta-beta/low gamma PSI of 1:3). For music, the dominance of PSIs 1:3 and 2:3 across genres and instruments suggests more tightly controlled rhythmic dependencies than for speech. 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 have arisen 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 6, the PSI between delta- and theta-rate AM bands was consistently the largest PSI in both the S-AMPH and PAD models. Again, this finding is consistent with our prior findings for IDS and rhythmic CDS [7,9]. As a further check, we also examined the PSI of sounds found in nature. The human hearing system has been receiving these quasi-rhythmic sounds at least as long as it has been receiving language and music, but unlike language and music, these sounds have not been produced by humans and shaped by human physiology and culture. Accordingly, it would not be expected that the temporal modulation structure of these natural sounds would be shared with IDS and CDS. The results showed that compared with music, the PSI between delta- and theta-rate AM bands was not consistently the largest PSI (S5 Appendix). This shows that the strong phase dependence between slower bands of AMs revealed for music and for IDS/CDS is not an artifact of the modelling approaches employed, but a core physical feature of their rhythmic structure. This shared structure is demonstrated in Fig 7, which depicts music, IDS, nature sounds (averaged) and a man-made rhythmic sound (a machine) from an envelope demodulation perspective. Comparison of the temporal structures of these sounds for the low-frequency modulation rates (0 –

5 Hz) shows that only music and speech show strong delta- and theta-AM band patterning. The nested structure of AM patterning across the higher modulation bands (12-40Hz) is also clearly visible for each quasi-rhythmic sound for all the natural sounds. This patterning is clearly absent for the man-made rhythmic sound of a machine.

Accordingly, the strong rhythmic character and acoustic temporal regularity of both infant- and child-directed speech and Western music appears to be influenced by AMs in the delta band (a 2 Hz modulation peak, in music reflecting a 120 bpm rate) and by delta-theta AM phase alignment. Our modelling data for temporal frequency (i.e., "*rhythm*") also map nicely to the Plomp and Levelt [70] modeling of the dissonance curve for spectral frequency (i.e., "*pitch*") (shown in Fig 6, Bottom). This may imply that these physical properties of fast spectral frequencies are also involved in very slow temporal modulation envelopes below 40 Hz.

Simulation Analyses

Finally, to investigate whether the detected (dissonance curve-like) characteristics really represented systematic features of natural musical rhythm, we conducted simulation analyses. Synthesized rhythmic but non-musical sounds that consisted of four temporal modulation envelopes (i.e., 2Hz, 4Hz, 8Hz and 16Hz modulators) and one spectral frequency (carrier) were summarized and the resulting compound tone waveform was then multiplied by 200 Hz sine

waves. The final synthesized waveform comprised a sound that included clear rhythmic information of delta (2Hz), theta (4Hz), alpha (8Hz) and gamma (16Hz) rhythms, and spectral information of a pitch around that of natural human voices (200 Hz) (for the figure, see S5 Appendix). The percept is similar to a harsh rhythmic whisper. As all of the temporal envelopes were comprised of simple sine waves with frequencies of a power of 2, PSI analyses of these artificial sounds should clearly and consistently reveal 1:2 integer ratios compared with other integer ratios. This was the case. Thus, the simulation analyses revealed that the PSIs for natural Western musical genres were different from those for artificial rhythmic sounds. This suggests that natural musical rhythm has covert and systematic integer ratios (i.e., 2:3, 3:4 and 4:5 as well as 1:2) within the AM hierarchy, at least when considering Western musical genres.

Discussion

Here we tested the prediction that the physical stimulus characteristics that describe IDS and CDS from a demodulation perspective would also describe the hierarchical rhythmic relationships that characterize music. Decomposition of the amplitude envelope of IDS and CDS has previously revealed that (a) the modulation peak in IDS is ~2 Hz [7], (b) that perceived rhythmic patterning depends on the three core AM bands in the amplitude envelope that are found systematically across the spectral range of speech [9], and (c) that varying metrical patterns such as trochaic and iambic meters can be identified by the phase relations between two of these bands of AMs (delta- and theta-rate AMs) [8]. The phase alignment (rhythmic synchronicity) of these relatively slow AM rates represents a unique statistical clue to rhythmic patterns [24]. Accordingly, we predicted that similar physical stimulus characteristics of the amplitude envelope of different musical genres and of music produced by different instruments would describe the underlying rhythmic structures.

Our demodulation perspective indeed revealed an hierarchy of temporal modulations that systematically described the acoustic properties of musical rhythm for a range of Western musical genres and instruments. The modelling indicated highly similar acoustic properties to IDS and CDS: a 2Hz modulation peak, particularly strong phase alignment between delta- and theta-rate AM bands across genres, and a distinct set of preferred PSIs that indicated multi-timescale synchronization across different AM bands. As the brain begins

learning language using IDS, and consolidates this learning via the rhythmic routines of the nursery (CDS), the present findings are consistent with the view that both music and language may (at least early in development, prior to acquiring expertise) rely on processing of the same physical stimulus characteristics. Although not tested directly here, it is likely that similar oscillatory entrainment mechanisms are used for encoding this hierarchical AM structure in both domains [43-46].

The multi-timescale synchronization found here was systematic across musical genres and instruments, suggesting that this AM hierarchy contributes to building perceived rhythmic structures. The nested AM hierarchies in music may yield nested musical units (crotchets, quavers, demiquavers and onsets), just as nested AM hierarchies in CDS yield linguistic units like syllables and rhymes [9]. The current modelling shows that acoustically-emergent musical units can in principle be parsed reliably from the temporal modulation spectra of the different musical genres examined, and that these units are reflected in each of delta-, theta-, alpha- and beta/gamma-rate bands of AM (Fig 8). To the best of our knowledge, our study is the first to reveal a shared hierarchical AM structure related to musical rhythm.

The modelling further revealed strong mutual dependence (using MI estimates) between adjacent bands in the AM hierarchy across musical genres (Western classical, jazz, rock, children's songs) and musical instruments (piano, guitar, violin, viola, cello, bass, single-voice, multi-voice). In particular, the mutual dependence between delta- and theta-rate bands of AM was the strongest

dependence identified by both models. Stronger mutual dependence between delta- and theta-rate AM bands characterises IDS and CDS, but could not be detected in non-music nature sounds that are quasi-rhythmic (e.g., fire, wind, rain, storms and rivers). The modelling suggests that for music, delta-theta phase alignment of AM bands may underpin metrical rhythmic patterns, at least for Western musical genres.

Convergent results from the phase synchronisation analyses further showed that multi-timescale synchronization between delta- and theta-rate AM bands was always higher than the other PSIs regardless of the integer ratios. This was not replicated for nature sounds. 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 [24]. For example, individual differences in sensitivity to delta- and theta- AM rates and their associated rise times are known to be implicated in disorders of language development [6,24]. In experimental studies, the phase alignment between the slower AM bands (delta and theta) in speech has been demonstrated to play a key role in the perception of metrical patterning (for example, judging whether tone-vocoded nursery rhymes were trochaic or iambic) [8]. Perceptual data from adults showed that a strong or stressed syllable was perceived when delta and theta modulation peaks were in alignment. The placement of stressed syllables governs metrical patterning in speech (e.g., trochaic, iambic and dactyl meters). The present findings concerning mutual dependence and phase synchronisation indicate that music may share these

properties: phase alignment between delta- and theta-rate AM bands may contribute to establishing musical metrical structure as well.

Accordingly, prior claims that the rhythmic properties of music and language are distinct [12], with the modulation spectrum for music peaking at 2 Hz and the modulation spectrum for speech peaking at 5 Hz, appear to arise from the exclusive reliance of the speech modelling on ADS. By contrast, our modelling approach shows better matching with temporal data from studies of IDS, where the modulation spectrum also peaks at 2 Hz. We would predict that the statistical regularities in temporal modulations may be the same for other forms of music, and for IDS and CDS in other languages, this remains to be explored. 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.

Another interesting result from the phase synchronisation analyses was the appearance of systematic 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 [1,71], and even the songs of different species [72]. For example, even prior to the acquisition of culture-specific biases of musical rhythm, young infants (5-month-olds) are influenced by ratio

complexity [73]. Our modelling further suggests that the AM bands in music 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 Fig 7 and the dissonance curve measured by Plomp & Levelt [70]). 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 [74]. Our modelling appears to suggest that the simple integer ratios (i.e., 1:2, 1:3, and 2:3) in the AM hierarchy comprise a fundamental set of statistics for musical rhythm perception. This fits well with prior data by [75], who demonstrated that certain integer ratios are prominent across music from both Western and non-Western cultures. The modelling suggests that AM phase hierarchies may play as strong a role as harmonicity regarding universal mechanisms of human hearing that are important for both music and language.

The modelling presented here is also relevant to the remediation of childhood language disorders. The possible utility of musical interventions for children with disorders of language learning such as developmental language disorder (DLD) and developmental dyslexia has long been recognised [76-80]. Such interventions are likely to be most beneficial when the temporal hierarchy of the music corresponds to the temporal hierarchy underpinning speech rhythm [6,24]. Careful consideration of the statistical rhythm structures characterizing speech in different languages may thus lead to better remedial outcomes. Similar

interventions could be beneficial for second language learners. A caveat is that here we modelled musical genres that could be designated WEIRD corpora (originating from Westernized, educated, industrialised, rich and democratic societies). Accordingly, further studies are necessary to understand how music interventions can contribute to improving speech processing in other languages.

In conclusion, the present study revealed that the physical stimulus characteristics that describe Western musical genres from an amplitude envelope decomposition perspective match those that describe IDS and CDS. The physical stimulus characteristics that describe ADS are different, suggesting that studies aiming to discover commonalities between music and speech should not rely exclusively on adult speech corpora. The modelling demonstrates a core acoustic hierarchy of amplitude modulations that yield musical rhythm across the amplitude envelopes of different Western musical genres and instruments, with mutual dependencies between AM bands playing a key role in organising rhythmic units in the musical hierarchy for each genre. Accordingly, biological mechanisms that exploit AM hierarchies may underpin the perception and development of both language and music. In terms of evolution, the novel acoustic statistics revealed here could also explain cross-cultural regularities in musical systems [71]; this remains to be tested.

Materials and Methods

The music samples for modelling consisted of the music corpora used in the study by Ding et al. [12], with the addition of 23 children's songs, in order to characterize more general properties of modulation spectra across musical genres. The final samples consisted of over 39 h of recordings (sampling rate = 44.1 kHz) of Western music (Western-classical music, Jazz, Rock, and Children's song) and musical instruments (single-voice: Violin, Viola, Cello, and Bass; multi-voice: Piano and Guitar). In addition, a range of natural sounds like wind and rain were extracted from sound files available on the internet (https://mixkit.co; https://www.zapsplat.com). 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: Spectral Amplitude Modulation Phase Hierarchy (S-AMPH, [8]), and 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 hierarchies in the amplitude envelopes of speech and music. 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 [22] (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 carrier or frequency modulation (FM) patterns [22,80]. 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 [22]. 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 resulting temporal modulation hierarchies.

Spectral Amplitude Modulation Phase Hierarchy (S-AMPH) model

Signal processing: spectral and temporal modulations. The methodologies were based on a previous study by Leong and Goswami [9]. To establish the patterns of spectral modulation, the raw acoustic signal was passed through a 28 log-spaced equivalent rectangular bandwidth (ERB_N) filterbank spanning 100–7250 Hz, which simulates the frequency decomposition by the cochlea in a normal human [18,65]. For further technical details of the filterbank design, see Stone and Moore [81]. 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 spectral patterning 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. To establish the patterns of 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 amplitude modulation hierarchy was defined by PCA. This can clarify co-activation patterns across modulation rate channels.

PCA to finding the core modulation hierarchy 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 hierarchy. 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 [9].

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):

$$y_t = m_t * c_t \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Probabilistic amplitude demodulation (PAD) [10] implements the amplitude demodulation as a problem of learning and inference. Learning corresponds to the estimation of the parameters that describe these distributional constraints such as the expected time-scale of variation of the modulator. Inference corresponds to the estimation of the modulator and carrier from the signals based on the learned or manually defined parametric distributional constraints. This information is encoded probabilistically in the likelihood: $P(y_{1:T}|c_{1:T}, m_{1:T}, \theta)$,

prior distribution over the carrier: $p(c_{1:T}|\theta)$, and prior distribution over the modulators: $p(m_{1:T}|\theta)$. Here, the notation $x_{1:T}$ represents all the samples of the signal x, running from 1 to a maximum value T. Each of these distributions depends on a set of parameters θ, which controls factors such as the typical time-scale of variation of the modulator or the frequency content of the carrier. For more detail, the parametrised joint probability of the signal, carrier and modulator is:

$$P(y_{1:T}, c_{1:T}, m_{1:T}|\theta) = P(y_{1:T}|c_{1:T} ;m_{1:T}, \theta) * p(c_{1:T}|\theta) * p(m_{1:T}|\theta) \qquad \ldots\ldots. (2)$$

Bayes' theorem is applied for inference, forming the posterior distribution over the modulators and carriers, given the signal:

$$P(c_{1:T}, m_{1:T}|y_{1:T}, \theta) = P(y_{1:T}, c_{1:T}, m_{1:T}|\theta) / P(y_{1:T}|\theta) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The full solution to PAD is a distribution over possible pairs of modulator and carrier. The most probable pair of modulator and carrier given the signal is returned:

$$m^*_{1:T}, c^*_{1:T} = \mathrm{argmax}\ P(c_{1:T}, m_{1:T}|y_{1:T}, \theta) \qquad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

That is, compared with S-AMPH, PAD does not implement the Hilbert transform, filtering, and PCA, but estimates the most appropriate modulator

(envelope) and carrier based on Bayesian inference that identifies the envelope which best matches the data and the a priori assumptions [10,22]. The solution takes the form of a probability distribution which describes how probable a particular setting 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 [22] (Fig 2). The positive slow envelope is modelled 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 samples, 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 over time. High modulation

frequency corresponds to fast modulations and vice versa (see Fig 2). The modulation spectra were averaged across all samples in each genre (Western-classical music, Jazz, Rock, and Children's song) and musical instruments (single-voice: Violin, Viola, Cello, and Bass; multi-voice: Piano and Guitar).

Mutual information between different modulation bands

We also examined how a tier of the temporal hierarchy of music may be mutually dependent on the timing of 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 [82].

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 adjacent-band oscillators are not independent

of, but interdependent on each other [52,63]. Given the evidence, we hypothesize that the adjacent tiers that can connect via so-called "*branches*" 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 applicability of an AM hierarchy to music as well as speech.

To explore 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) were 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 the 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, l, g, e in Fig 2). Then, using the phase angle values, the mutual information between different temporal modulation bands was measured.

Phase Synchronization Analyses

Based on the findings of mutual information, we further investigated possible multi-timescale phase synchronization between bands by computing the integer ratios between "*adjacent*" AM hierarchies (i.e., the number of parent vs. daughter elements in an AM hierarchy). This analysis addresses how many daughter elements a parent element encompasses in general in a particular musical genre. 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, l, 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; [68]), and was subsequently adapted for neural analyses of oscillatory phase-locking [69]. 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\

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 and systematic integer ratios in an AM hierarchy.

Acknowledgments

We thank Jose Pérez-Navarro (Basque Center on Cognition, Brain and Language) for helpful discussion and Eri Daikoku (a Western-classical pianist) for providing useful advice to generate Fig 8.

Competing interests

The authors declare no competing financial interests.

References

1. Mehr, S., Krasnow, M., Bryant, G. A. & Hagen, E. Origins of music in credible signaling. *Behavioral and Brain Sciences*. 2020; 1-41. doi: 10.1017/S0140525X20000345
2. Patel, A. D. (2008). *Music, language, and the brain*. Oxford: Oxford University Press.
3. Falk, D. (2004). Prelinguistic evolution in early hominims: Whence Motherese? *Behavioural and Brain Sciences, 27*, 491-541.
4. Mehler, J. et al. A precursor of language acquisition in young infants. *Cognition.* 1988;**29**:143–178.
5. Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences, 15 (1*), 3-10.
6. Goswami, U. (2019). A Neural Oscillations Perspective on Phonological Development and Phonological Processing in Developmental Dyslexia. *Linguistic Compass, 13,* e12328
7. Leong, V., Kalashnikova, M., Burnham, D. & Goswami, U. The Temporal Modulation Structure of Infant-Directed Speech. *Open Mind.* 2017;**1**:78–90. doi:10.1162/OPMI_a_00008
8. Leong, V., Stone, M. A., Turner, R. E. & Goswami, U. A role for amplitude modulation phase relationships in speech rhythm perception. *J Acoust Soc Am.* 2014;**136**:366–381.

9. Leong, V. & Goswami, U. Acoustic-Emergent Phonology in the Amplitude Envelope of Child-Directed Speech. *PLoS One.* 2015;**10**:e0144411. Available: https://doi.org/10.1371/journal.pone.0144411

10. Turner, R. E. & Sahani, M. Demodulation as Probabilistic Inference. I*EEE Trans Audio Speech Lang Processing.* 2011;**19**:2398–2411. doi:10.1109/TASL.2011.2135852

11. Araújo, J., Flanagan, S., Castro-Caldas, A. & Goswami, U. The temporal modulation structure of illiterate versus literate adult speech. *PLoS One.* 2018;**13**:e0205224.

12. Ding, N. et al. Temporal modulations in speech and music. *Neurosci Biobehav Rev.* 2017;**81**:181–187. doi:https://doi.org/10.1016/j.neubiorev.2017.02.011

13. Trehub, S. E., & Trainor, L. (1998). Singing to infants: Lullabies and play songs. In C. Rovee-Collier, L. Lipsitt & H. Hayne (Eds.), *Advances in infancy research (Vol. 12)* (pp. 43-77). Ablex.

14. Hayes, B. Metrical stress theory: Principles and case studies. *University of Chicago Press*; 1995.

15. Tierney, A. & Kraus, N. Music training for the development of reading skills. *Prog Brain Res.* 2013;**207**:209–241. doi:10.1016/B978-0-444-63327-9.00008-4

16. Schön, D. et al. Similar cerebral networks in language, music and song perception. *Neuroimage.* 2010;**51**:450–461. doi:https://doi.org/10.1016/j.neuroimage.2010.02.023

17. Peretz, I., Vuvan, D., Lagrois, M-É. & Armony, J. L. Neural overlap in processing music and speech. *Philos Trans R Soc B Biol Sci.* 2015;**370**: 20140090. doi:10.1098/rstb.2014.0090

18. Moore, B. C. J. An introduction to the psychology of hearing. *Brill*; 2012.

19. Moore, B. C. J. (**2019**). "The roles of temporal envelope and fine structure information in auditory perception," Acoust. Sci. Tech. **40**, 61-83.

20. Glasberg, B. R., Moore, B. C. J. Derivation of auditory filter shapes from notched-noise data. *Hear Res.* 1990*;***47**:103–138. doi:https://doi.org/10.1016/0378-5955(90)90170-T

21. Leong, V. Prosodic rhythm in the speech amplitude envelope: Amplitude modulation phase hierarchies (AMPHs) and AMPH models. *Cambridge, UK PhD Thesis*. 2012.

22. Turner, R. Statistical models for natural sounds. Dr Diss Coll London. 2010.

23. Bhide, A., Power, A. & Goswami, U. A Rhythmic Musical Intervention for Poor Readers: A Comparison of Efficacy With a Letter-Based Intervention. *Mind, Brain, Educ.* 2013;**7**:113–123. doi:10.1111/mbe.12016

24. Goswami, U. Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Ann N Y Acad Sci.* 2019;**1453**:67–78.

25. Morton, J., Marcus, S. & Frankish, C. Perceptual centers (P-centers). *Psychol Rev.* 1976;**83**:405.

26. Hoequist, Jr C. E. The perceptual center and rhythm categories. *Lang Speech.* 1983;**26**:367–376.

27. Scott, S. P-centres in speech-an acoustic analysis. *PhD thesis, Univ Coll London*. 1993.

28. Gordon, J. W. The perceptual attack time of musical tones. *J Acoust Soc Am.* 1987;**82**:88–105.

29. Huss, M., Verney, J. P., Fosker, T., Mead, N. & Goswami, U. Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex.* 2011;**47**:674–689. doi:10.1016/j.cortex.2010.07.010

30. Goswami, U., Huss, M., Mead, N., Fosker, T., & Verney, J. (2013). Perception of patterns of musical beat distribution in phonological developmental dyslexia: Significant longitudinal relations with word reading and reading comprehension. *Cortex, 49 (5),* 1363-1376.

31. Doelling, K. B., Arnal, L. H., Ghitza, O. & Poeppel, D. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage.* 2014;**85**:761–768.

32. Liberman, M. & Prince, A. On stress and linguistic rhythm. *Linguist Inq.* 1977;**8**: 249–336.

33. Selkirk, E. Phonology and Syntax. The Relation between Sound and Structure. *Cambridge, Mass*; 1984.

34. Selkirk, E. O. The Role of Prosodic Categories in English Word Stress. *Linguist Inq.* 1980;**11**:563–605. Available: http://www.jstor.org/stable/4178179

35. Lerdahl, F., Jackendoff, R. & Jackendoff, R. S. A Generative Theory of Tonal Music. *MIT Press*; 1983. Available: https://books.google.de/books?id=38YcngEACAAJ

36. Luo, H. & Poeppel, D. Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron.* 2007;**54**:1001–1010. doi:https://doi.org/10.1016/j.neuron.2007.06.004

37. Ahissar, E. et al. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci.* 2001;**98**:13367 LP-13372. doi:10.1073/pnas.201400998

38. Giraud, A. L. & Poeppel, D. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat Neurosci.* 2012*;***15**:511–517. doi:10.1038/nn.3063

39. Henry, M. J. & Obleser, J. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc Natl Acad Sci.* 2012**;109**:20095–20100.

40. Overath, T., McDermott, J. H., Zarate, J. M. & Poeppel, D. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci.* 2015;**18**:903.

41. Ding, N., Melloni, L., Zhang, H., Tian, X. & Poeppel, D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci.* 2016*;***19**:158–164. doi:10.1038/nn.4186

42. Park, H., Ince, R. A. A., Schyns, P. G., Thut, G. & Gross, J. Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to

Continuous Speech in Human Listeners. *Curr Biol.* 2015;**25**:1649–1653. doi:10.1016/j.cub.2015.04.049

43. Doelling, K. B. & Poeppel, D. Cortical entrainment to music and its modulation by expertise. *Proc Natl Acad Sci.* 2015;**112**:E6233–E6242.

44. Norman-Haignere, S., Kanwisher, N. G. & McDermott, J. H. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron.* 2015;**88**:1281–1296.

45. Nozaradan, S., Peretz, I., Missal, M. & Mouraux, A. Tagging the neuronal entrainment to beat and meter. *J Neurosci.* 2011;**31**:10234–10240.

46. Baltzell, L. S., Srinivasan, R. & Richards, V. Hierarchical organization of melodic sequences is encoded by cortical entrainment. *Neuroimage.* 2019;**200**:490–500. doi:https://doi.org/10.1016/j.neuroimage.2019.06.054

47. Coath, M. et al. (2010). Model cortical responses for the detection of perceptual onsets and beat tracking in singing. Connection Science, 21:2-3, 193-205, DOI:10.1080/09540090902733905

48. Large, EW, CS Wasserman, CS, E Skoe, E, HL Read HL (2019). Neural entrainment to missing pulse rhythms. *The Journal of the Acoustical Society of America* 144 (3), 1760-1760.

49. Large, E.W., & Jones, M.R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106*, 1-119.

50. Poeppel, D. The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time.’ *Speech*

Commun. 2003;**41**:245–255. doi:https://doi.org/10.1016/S0167-6393(02)00107-3

51. Obleser, J. & Kayser, C. Neural Entrainment and Attentional Selection in the Listening Brain. *Trends Cogn Sci.* 2019;**23**:913–926. doi:10.1016/j.tics.2019.08.004

52. Gross, J et al. Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLOS Biol.* 2013;**11**:e1001752. Available: https://doi.org/10.1371/journal.pbio.1001752

53. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr Biol.* 2015;**25**:2457–2465.

54. Harding, E.E., Sammler, D., Henry, M.J., Large, E.W., & Kotz, S.A. (2019). Cortical tracking of rhythm in music and speech. *Neuroimage, 185*, 96-101.

55. Greenberg, S. (2006). A multi-tier framework for understanding spoken language. In S. Greenberg & W. Ainsworth (Eds.), *Listening to speech: An auditory perspective*. Lawrence Erlbaum Associates.

56. Dau, T., Kollmeier, B. & Kohlrausch, A. Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *J Acoust Soc Am.* 1997*;***102**:2906–2919.

57. Yang, X., Wang, K. & Shamma, S. A. Auditory representations of acoustic signals. *IEEE Trans Inf theory.* 1992;**38**:824–839.

58. Joris, P. X., Schreiner, C. E. & Rees, A. Neural processing of amplitude-modulated sounds. *Physiol Rev.* 2004;**84**:541–577.

59. Shamma, S. On the role of space and time in auditory processing. *Trends Cogn Sci.* 2001;**5**:340–348.

60. Barton, B., Venezia, J. H., Saberi, K., Hickok, G. & Brewer, A. A. Orthogonal acoustic dimensions define auditory field maps in human cortex. *Proc Natl Acad Sci.* 2012;**109**:20738–20743.

61. Santoro, R. et al. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput Biol.* 2014;**10**:e1003412.

62. Fontolan, L., Morillon, B., Liegeois-Chauvel, C. & Giraud, A. L. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat Commun.* 2014;**5**:1–10. doi:10.1038/ncomms5694

63. Lakatos, P. et al. An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *J Neurophysiol.* 2005;**94**:1904–1911. doi:10.1152/jn.00263.2005

64. Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C., Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. USA **102**, 2293-2298. https://doi.org/10.1073/pnas.0406460102.

65. Dau, T. & Kollmeir, B. (1997). Modelling auditory processing of amplitude modulation I. Detection and masking with narrowband carriers. *Journal of the Acoustical Society of America, 102*, 2892. https://doi.org/10.1121/1.420344

66. Klein W, Plomp R, Pols LCW. Vowel Spectra, Vowel Spaces, and Vowel Identification. *J Acoust Soc Am.* 1970;**48**:999–1009. doi:10.1121/1.1912239

67. Pols, L. C. W., Tromp, H. R. C. & Plomp, R. Frequency analysis of Dutch vowels from 50 male speakers. *J Acoust Soc Am.* 1973;**53**:1093–1101. doi:10.1121/1.1913429

68. Tass, P. et al. Detection of $\mathit{n}:\mathit{m}$ Phase Locking from Noisy Data: Application to Magnetoencephalography. *Phys Rev Lett.* 1998;**81**:3291–3294. doi:10.1103/PhysRevLett.81.3291

69. Schack, B. & Weiss, S. Quantification of phase synchronization phenomena and their importance for verbal memory processes. *Biol Cybern.* 2005;**92**:275–287. doi:10.1007/s00422-005-0555-1

70. Plomp, R. & Levelt, W. J. M. Tonal consonance and critical bandwidth. *J Acoust Soc Am.* 1965;**38**:548–560.

71. McPherson, M. J. et al. Perceptual fusion of musical notes by native Amazonians suggests universal representations of musical intervals. *Nat Commun.* 2020;**11**:2786. doi:10.1038/s41467-020-16448-6

72. Roeske, T. C., Tchernichovski, O., Poeppel, D. & Jacoby, N. Categorical rhythms are shared between songbirds and humans. *Current Biology.* 2020;**30**:3544-55.

73. Hannon, E.E., Soley, G. and Levine, R.S. (2011), Constraints on infants' musical rhythm perception: effects of interval ratio complexity and enculturation. *Developmental Science, 14*: 865-872.https://doi.org/10.1111/j.1467-7687.2011.01036.x

74. Schwartz, D. A., Howe, C. Q. & Purves, D. The statistical structure of human speech sounds predicts musical universals. *J Neurosci.* 2003;**23**:7160–7168.

75. Jacoby, N., & McDermott, J.H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology, 27* (3), 359-70.

76. Ladányi, E., Persici, V., Fiveash, A., Tillmann, B. & Gordon, R. L. Is atypical rhythm a risk factor for developmental speech and language disorders? *Wiley Interdiscip Rev Cogn Sci.* 2020;e1528.

77. Cumming, R., Wilson, A., Leong, V., Colling, L. J. & Goswami, U. Awareness of Rhythm Patterns in Speech and Music in Children with Specific Language Impairments. *Front Hum Neurosci.* 2015;**9**:672. doi:10.3389/fnhum.2015.00672

78. Kodály, Z. The Selected Writings of Zolta´n Koda´ly. *London: Boosey and Hawkes, [Lily Hala´py and Fred Macnicol, Trans.]*. 1974.

79. Jacques-Dalcroze, E. Rhythm, Music and Education [*Rubinstein H, Trans.]. London: The Dalcroze Society Inc*. 1980.

80. Elliott, T. M. & Theunissen, F. E. The Modulation Transfer Function for Speech Intelligibility. *PLOS Comput Biol.* 2009;**5**:e1000302. Available: https://doi.org/10.1371/journal.pcbi.1000302

81. Stone, M. A. & Moore, B. C. J. Tolerable hearing aid delays. III. Effects on speech production and perception of across-frequency variation in delay. *Ear Hear.* 2003;**24**:175–183.

82. Daikoku, T. Entropy, Uncertainty , and the Depth of Implicit Knowledge on Musical Creativity□: Computational Study of Improvisation in Melody and Rhythm. *Front. Comput. Neurosci.* 2018;**12**:1–11. doi:10.3389/fncom.2018.00097

Supporting information

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.

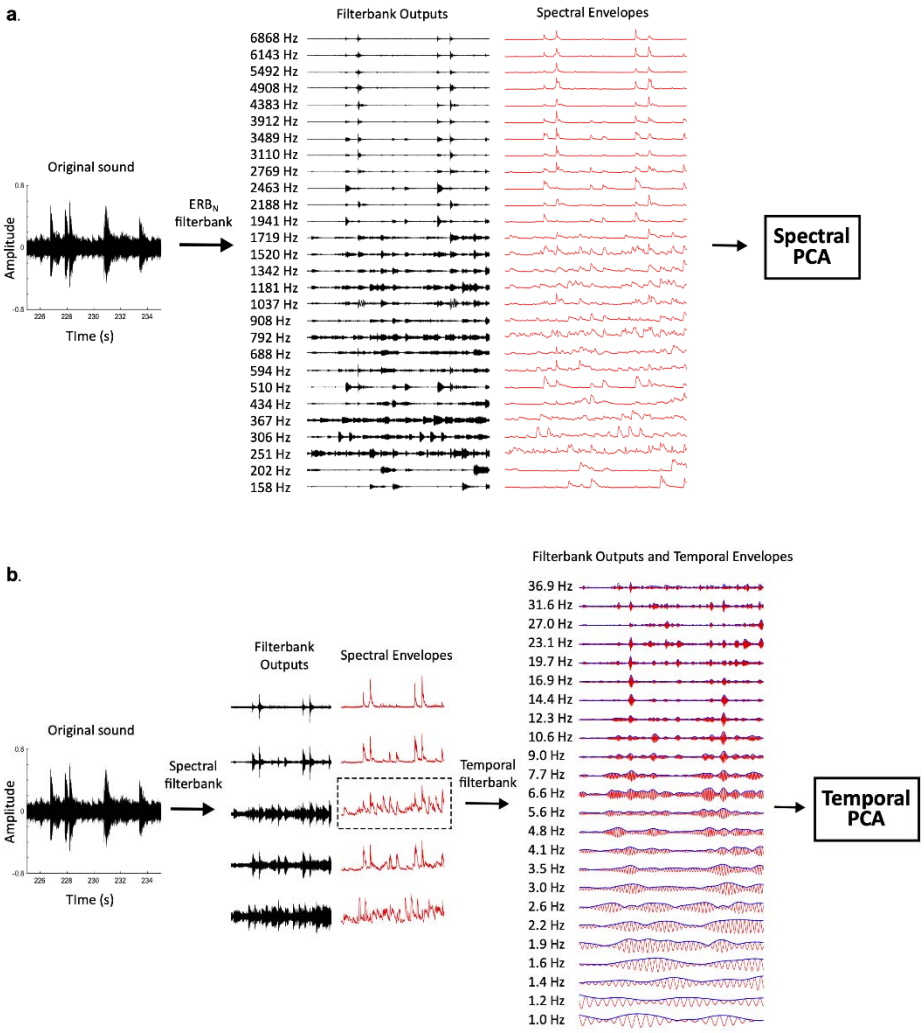
a.
Original sound
Amplitude
Time (s)
ERB$_N$ filterbank
Filterbank Outputs
Spectral Envelopes
6868 Hz
6143 Hz
5492 Hz
4908 Hz
4383 Hz
3912 Hz
3489 Hz
3110 Hz
2769 Hz
2463 Hz
2188 Hz
1941 Hz
1719 Hz
1520 Hz
1342 Hz
1181 Hz
1037 Hz
908 Hz
792 Hz
688 Hz
594 Hz
510 Hz
434 Hz
367 Hz
306 Hz
251 Hz
202 Hz
158 Hz
Spectral PCA
b.
Original sound
Amplitude
Time (s)
Spectral filterbank
Filterbank Outputs
Spectral Envelopes
Temporal filterbank
Filterbank Outputs and Temporal Envelopes
36.9 Hz
31.6 Hz
27.0 Hz
23.1 Hz
19.7 Hz
16.9 Hz
14.4 Hz
12.3 Hz
10.6 Hz
9.0 Hz
7.7 Hz
6.6 Hz
5.6 Hz
4.8 Hz
4.1 Hz
3.5 Hz
3.0 Hz
2.6 Hz
2.2 Hz
1.9 Hz
1.6 Hz
1.4 Hz
1.2 Hz
1.0 Hz
Temporal PCA

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.

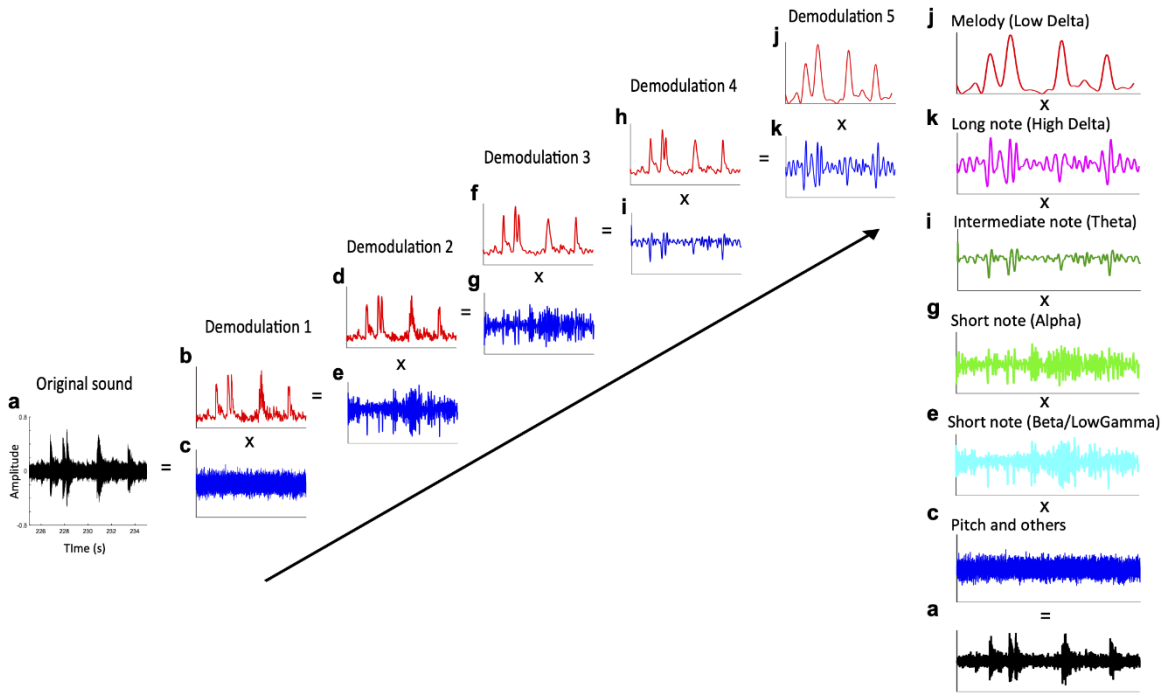

Fig 2. Signal processing steps in PAD model. Example of an amplitude modulation (AM) hierarchy derived by recursive application of PAD. In the first demodulation round (left column), the data (**a**) are demodulated using PAD set to a fast time scale. This yields a relatively quickly-varying envelope (**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**).

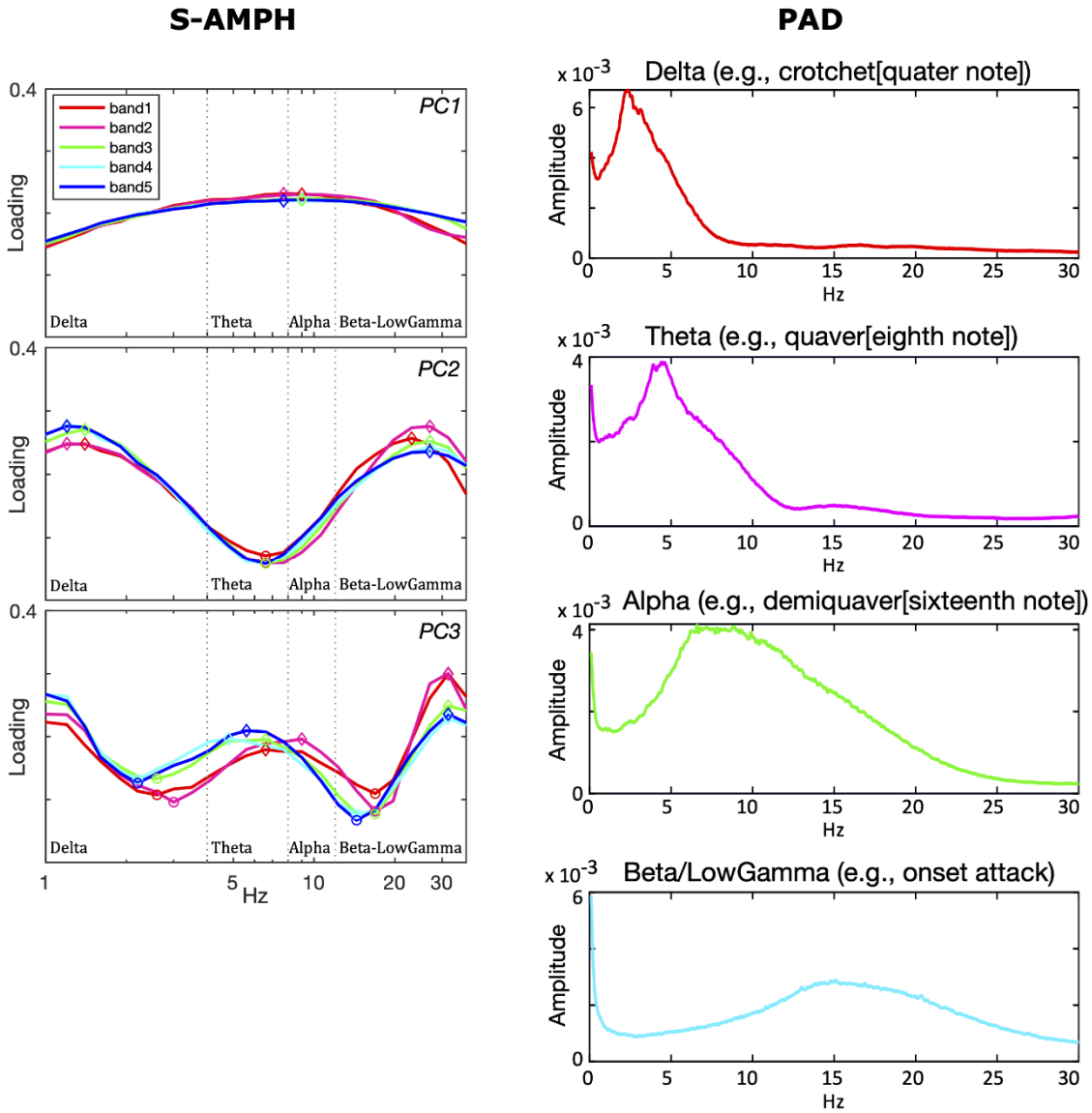

Fig 3. Core temporal modulation rates. Grand average absolute value of temporal PCA component loading patterns in the S-AMPH (a) model and modulation spectra of FFT in the PAD model (b). Both models showed an amplitude modulations (AM) hierarchy that consisted of delta-, theta-, alpha- and beta-rate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

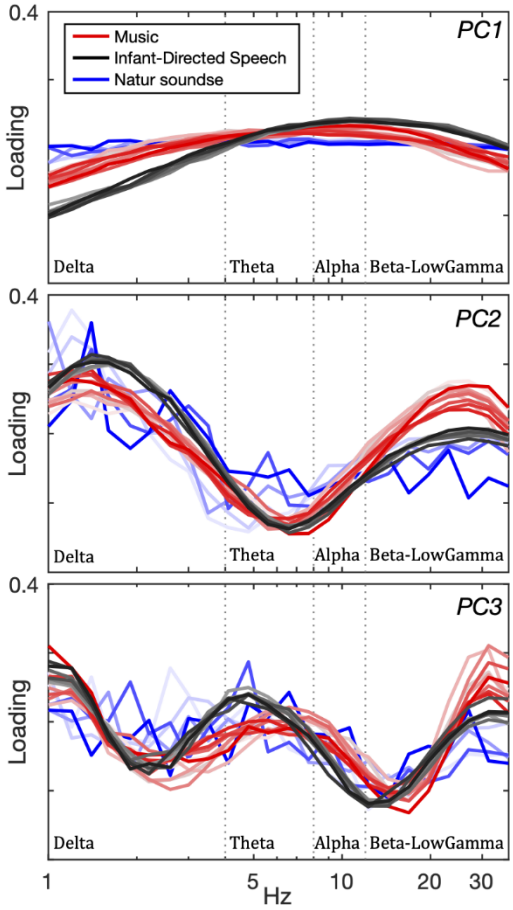

Fig 4. Core temporal modulation rates of music, speech, and nature sounds. Grand average absolute value of temporal PCA component loading patterns in the S-AMPH. Red, grey, and blue colour scales represent music, speech, and nature sounds, respectively. Individual lines represent different speakers, musical genres and nature sounds.

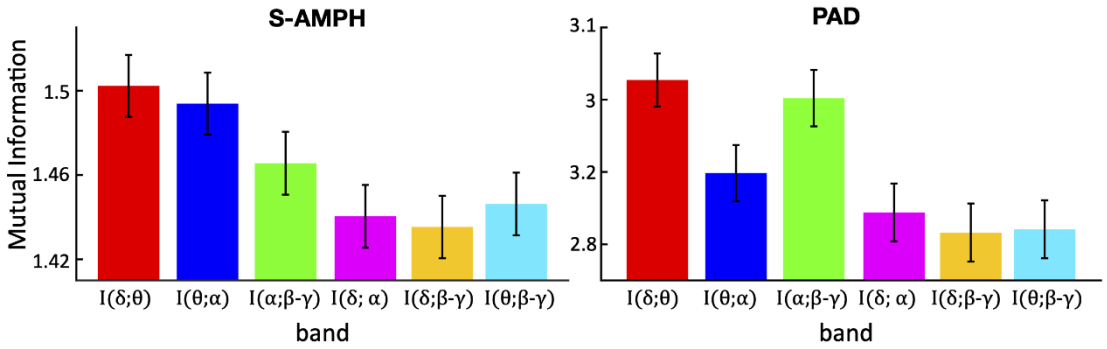

Fig 5. Mutual information between different tiers in an amplitude modulation **hierarchy.** Both S-AMPH (a) and PAD models (b) showed stronger dependency between each adjacent tier of an amplitude modulations (AM) hierarchy. Further, mutual dependence between delta- and theta-rate AM bands was the strongest of all mutual dependence in both models.

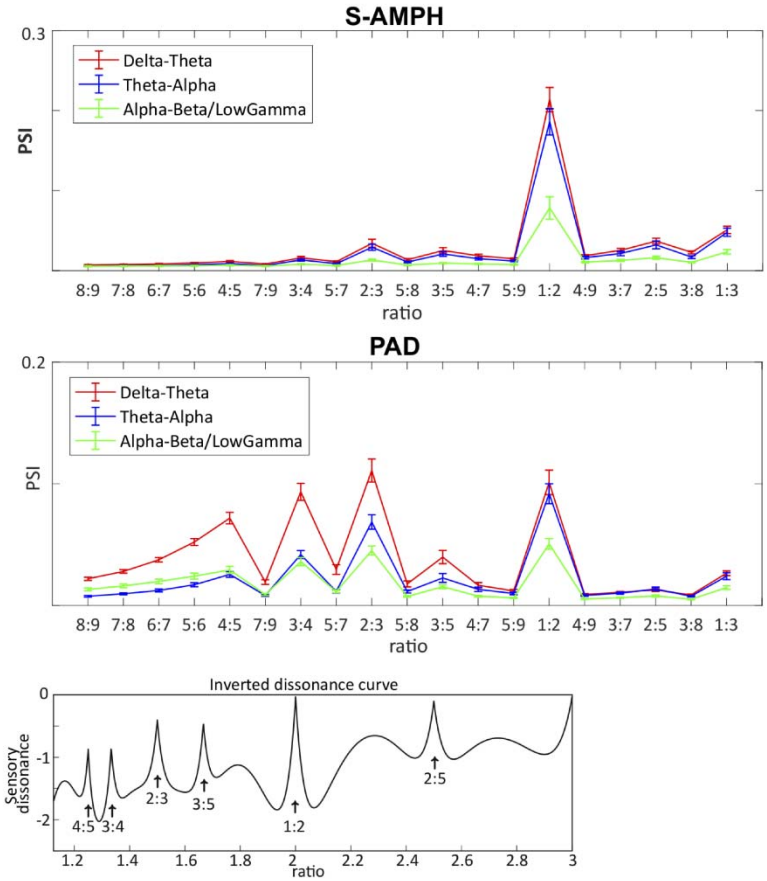

Fig 6. Phase synchronization index between different tiers in an amplitude modulation hierarchy. Both S-AMPH (a) and PAD models (b) showed that the simpler integer ratios (i.e., m/n) synchronize their phase with each other. The inverted dissonance curve **(c)** was obtained by including the first five upper partials of tones with a 440 Hz (i.e., pitch standard, A4) fundamental frequency in calculating the total dissonance of intervals [70]. It is of note that the peaks of PSI (a) correspond to those of the dissonance curve.

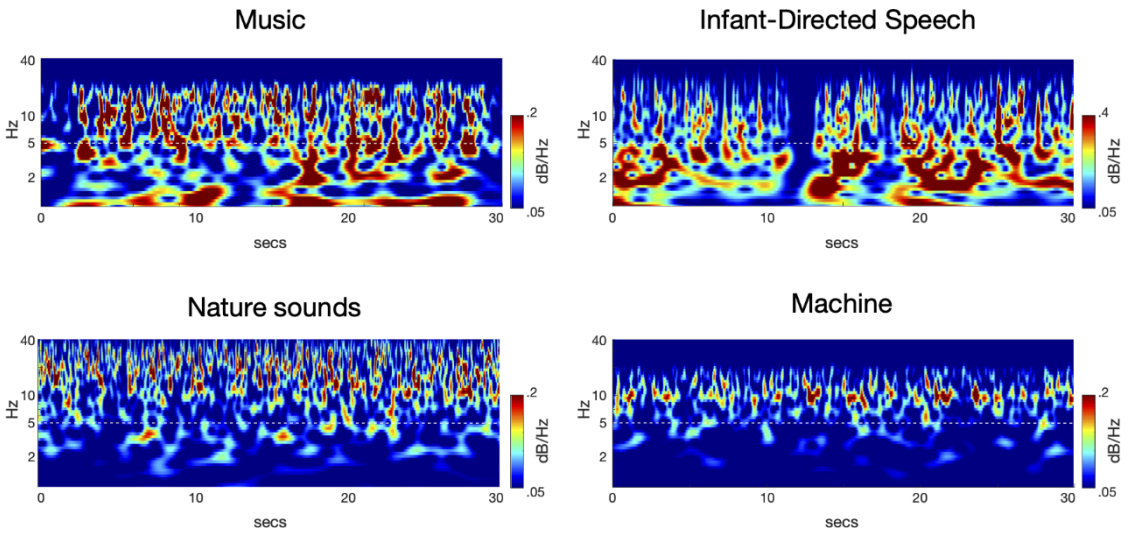

Fig 7. Scalograms depicting the amplitude modulation (AM) envelopes derived by recursive application of PAD (i.e., carrier). Continuous Wavelet Transform (CWT) was run on each AM envelope from randomly chosen 30-s excerpts of music, speech, nature sounds, and machine sounds. The x-axis denotes time (30 s) and the y-axis denotes modulation rate (0.1-40Hz). The maximal amplitude is normalized to 0 dB. It should be noted the low frequency structure (<5 Hz) visible in music and IDS is absent for the nature and machine sounds.

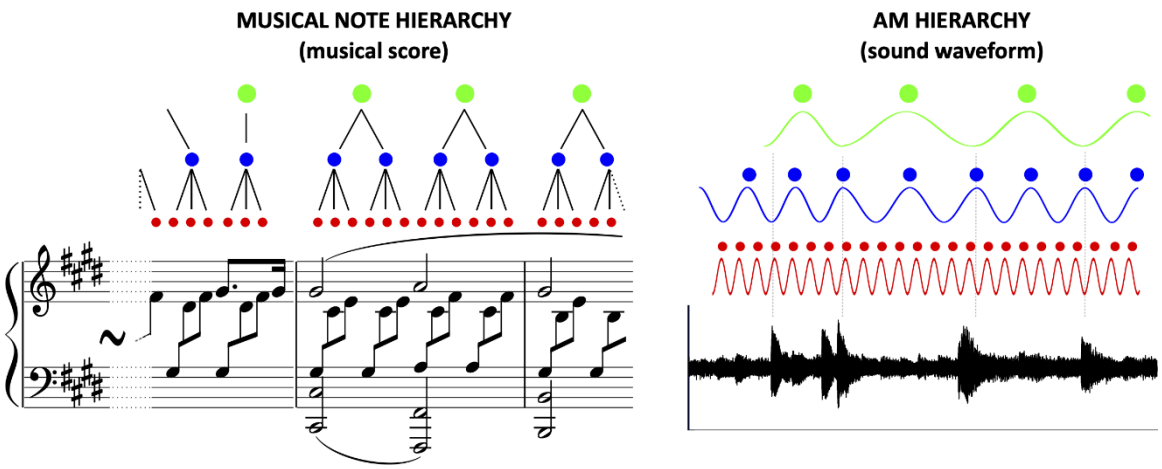

Fig 8. Hierarchical structure of rhythm in music. Left and right are the representation by musical score and the corresponding sound waveform of a part of the 33 Variations on a waltz by Anton Diabelli, Op. 120 (commonly known as the Diabelli Variations) by Ludwig van Beethoven. The musical rhythm could be hierarchically organised based either on note values (left) or amplitude modulations (AM, right). 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).