INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Kello CT, Bella SD, Médé B, Balasubramaniam R. 2017 Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. *J. R. Soc. Interface* **14**: 20170231.
http://dx.doi.org/10.1098/rsif.2017.0231

Received: 28 March 2017
Accepted: 12 September 2017

Subject Category:
Life Sciences–Physics interface

Subject Areas:
biocomplexity

Keywords:
hierarchical temporal structure, nested event clustering, speech, music, animal vocalizations

Author for correspondence:
Christopher T. Kello
e-mail: ckello@ucmerced.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3887803.

THE ROYAL SOCIETY PUBLISHING

Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes

Christopher T. Kello[1], Simone Dalla Bella[2,3,4,5], Butovens Médé[1] and Ramesh Balasubramaniam[1]

[1]Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA
[2]EuroMov Laboratory, Université de Montpellier, 700 Avenue du Pic Saint-Loup, 34090 Montpellier, France
[3]Institut Universitaire de France, 1 Rue Descartes, 75231 Paris, France
[4]International Laboratory for Brain, Music and Sound Research (BRAMS), 1430 Boulevard du Mont-Royal, Montreal, Quebec, Canada H2 V 2J2
[5]Department of Cognitive Psychology, WSFiZ in Warsaw, 55 Pawia Street, 01-030 Warsaw, Poland

CTK, 0000-0003-1588-9474

Humans talk, sing and play music. Some species of birds and whales sing long and complex songs. All these behaviours and sounds exhibit hierarchical structure—syllables and notes are positioned within words and musical phrases, words and motives in sentences and musical phrases, and so on. We developed a new method to measure and compare hierarchical temporal structures in speech, song and music. The method identifies temporal events as peaks in the sound amplitude envelope, and quantifies event clustering across a range of timescales using Allan factor (AF) variance. AF variances were analysed and compared for over 200 different recordings from more than 16 different categories of signals, including recordings of speech in different contexts and languages, musical compositions and performances from different genres. Non-human vocalizations from two bird species and two types of marine mammals were also analysed for comparison. The resulting patterns of AF variance across timescales were distinct to each of four natural categories of complex sound: speech, popular music, classical music and complex animal vocalizations. Comparisons within and across categories indicated that nested clustering in longer timescales was more prominent when prosodic variation was greater, and when sounds came from interactions among individuals, including interactions between speakers, musicians, and even killer whales. Nested clustering also was more prominent for music compared with speech, and reflected beat structure for popular music and self-similarity across timescales for classical music. In summary, hierarchical temporal structures reflect the behavioural and social processes underlying complex vocalizations and musical performances.

1. Introduction

Humans and other animals produce complex acoustic signals for various purposes. Speech, song and music serve a variety of functions including communication and entertainment. Long and varied vocalizations in certain whale and bird species are used for courtship, territorial establishment and social affiliation [1]. Intuitively, these vocalizations sound complex because they resemble human speech, song and music. Human signals are complex also in terms of their underlying syntax and meaning, which may be partly true of some whale songs and bird songs as well [2,3]. But syntax and meaning

aside, the present study is focused on complexities in the sounds themselves, and how they are expressed in terms of hierarchical temporal structure.

Currently there is no formalized method for quantifying and relating the complex sounds of speech, music, whale song and bird song. Most research has focused instead on analyses specific to each given type of signal. For instance, cues to timbre and rhythm are relevant to music [4], cues to prosody are relevant to speech [5], and sinusoidal pulses are relevant to bird calls [6]. However, one property that is common to all complex vocal and musical signals is their hierarchical temporal structure [7,8]. Speech production has articulatory features like plosivity that unfold over millisecond timescales, syllabic features that unfold over longer timescales, lexical and phrasal features over even longer timescales, and so on [9]. Likewise, musical performances have notes played over milliseconds, motives played over hundreds of milliseconds, phrases played over seconds, and so on [8]. Hierarchies of production units spanning embedded timescales also have been hypothesized for whale song and bird song [10,11], albeit we do not have the same firsthand insight into these units compared with speech and music.

Metrical theories of temporal hierarchies are well known in both speech [12] and music [13] research. We also posit temporal hierarchies as a useful basis for quantifying and comparing the complex structures of speech and music, and bird song and whale song as well. However, rather than posit abstract units and levels of a temporal hierarchy, we hypothesize that hierarchical temporal structure in complex vocalizations and musical performances is measurably reflected in the hierarchical temporal structure of their sound signals. We formulate a measure of temporal sound structure aimed at categorizing and comparing the coarse-grained shapes of temporal hierarchies across domains. We investigate how these shapes inform the behavioural and social processes that generate them.

Our measure of sound structure is based on instantaneous events in acoustic waveforms defined only by their moments in time. Each event is an acoustic feature of the sound signal, but clusters and other aggregations of events may relate to perceptual, motoric, behavioural and social processes involved in producing music and complex vocalizations. Transforming the waveform into a series of events has three main benefits. First and foremost, the transformation distils temporal information while leaving behind variability that is irrelevant to hierarchical temporal structure. Second, it normalizes temporal information in series of binary events that are directly comparable across different sound signals. Third, it results in a point process that is amenable to well-developed methods for quantifying hierarchical temporal structure.

There are numerous possible events that may be extracted from complex acoustic signals, but we chose peaks in the amplitude envelope because they are universal and simple to identify in the signal. Clusters of peaks, which are our basic units of analysis, reflect periods of relatively greater activity in acoustic energy for a given segment of time. The smallest clusters can group together to form larger clusters of clusters over longer periods of time, and so on, to form hierarchical temporal structures [14]. Note that our method does not require a one-to-one relationship between event clusters and units of perception or production, such as syllables or chords, although there will be correspondences with such units in many cases. Instead, we compute the overall degree of clustering, based on variance in the overall grouping of events, that is related to hierarchically nested units of vocalization and music. In support of this aim, a recent study [15] showed that variance in the clusters of peak events in speech is highly correlated with variance in phonetically transcribed durations of linguistic units across hierarchical levels like syllables, words, phrases and sentences. This result supports a link between event clustering and linguistic units at the aggregate level of variability across timescales. It does not speak to relationships at the level of individual clusters and units. Here, we take a similar approach by testing whether the clustering of events relates to domain-general dimensions of vocalization and music that shed light on their hierarchical temporal structures and relations between them.

Once the peak events and their times of occurrence are identified, we can use Allan factor (AF) analysis [16] to quantify the clustering of events in terms of their variances in timing at different timescales (see figure 1 for equation and illustration). Details are provided below, but in brief: windows of a given size are tiled across a time series of events, and events are counted within each window. The mean difference in event counts between adjacent windows is the basis for measuring clustering—differences increase as events cluster more within some windows and less within adjacent windows. The mean difference is normalized and computed for each of several different window sizes to yield a measure of clustering as a function of timescale. Clustering is hierarchical when smaller clusters in smaller windows combine to form larger clusters in larger windows, and larger clusters combine to form still larger clusters, and so on. The pattern of hierarchical clustering is captured by changes in AF variance as a function of timescale.

AF analysis and similar methods have been used in prior studies of speech event series [15,17–19]. Together, these studies showed that hierarchical temporal structure is present in speech at multiple timescales. Abney *et al*. [17] reported evidence that speakers in a conversation adapt the hierarchical temporal structure of their voices to each other, and argumentative conversations have more structure compared with affiliative conversations. Luque *et al*. [18] showed that hierarchical temporal structure follows a power law in fast timescales, and can be attributed to physiological processes of speech that are common to several different languages. Falk & Kello [15] showed that hierarchical temporal structure is greater in infant-directed speech compared with adult-directed speech, in conjunction with its relation to linguistic units as noted above.

In the present study, we find that all complex acoustic signals analysed are well-characterized by nested event clustering. Specific categories of signals are found to be associated with specific patterns of clustering across timescales. The overall result is a natural taxonomy of complex sounds in terms of the behavioural and social processes that generated them. The taxonomy illuminates some of the similarities, differences, and relationships underlying bird song, whale song, human speech and music.

2. Material and methods

We analysed four main categories of complex vocalizations and musical performances, each with four subcategories: (i) bird

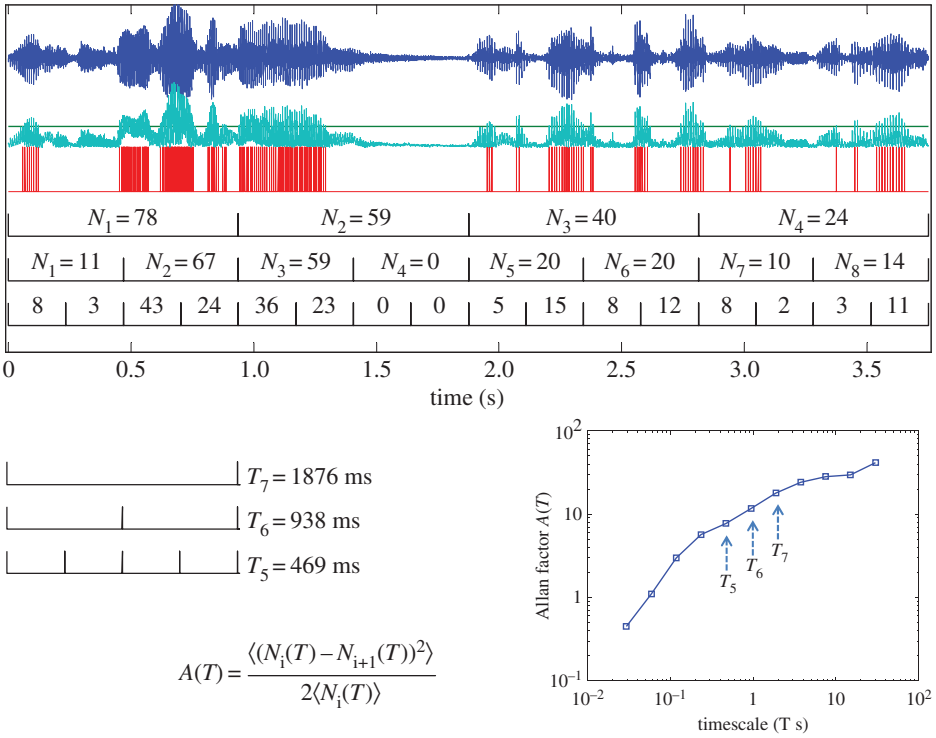

Figure 1. Illustration of Allan factor analysis using an example waveform segment from Michael Steger's TED talk entitled 'What makes life meaningful'. The waveform is at top (blue), followed by the Hilbert envelope (cyan) and the peak event series (red). Event counts N are shown inside brackets representing tiled windows for three different sizes T (where timescale = $2T$). The AF variance equation is also shown, along with the entire $A(T)$ function for this talk.

song and whale song, including nightingale, hermit thrush, humpback and killer whales (orca); (ii) human vocalizations, including original and synthesized TED talks, naturalistic interviews, and singing; (iii) popular music, including rock, instrumental pop/R&B, acappella, rap and electronic dance music; and (iv) classical music, including piano solos, violin concertos, guitar solos and symphonies. We compared these vocal and musical signals with jazz instrumentals as an additional category, and thunderstorm recordings as a complex but non-biological reference signal.

Ten recordings were chosen for each of 23 subcategories (see figure 3 and electronic supplementary material), except for thunderstorms for which there were only three recordings. Musical recordings were chosen to be prototypical of their genre, and TED talks were chosen based on popularity and the absence of sounds other than talking by the speaker. Conversational interviews were sampled randomly from 10 different speakers in the Buckeye speech corpus [20]. Recordings were downsampled from 44.1 KHz to 11 KHz to reduce the amount of data to be analysed (preliminary results showed no differences using the original sample rate). For stereo recordings, each channel was analysed separately and the resulting AF variances were averaged together. Recordings were chosen to be at least 4 min long, and window sizes were varied from approximately 15 ms to 15 s. Preliminary results showed that there was no need for windows shorter than 15 ms because events stopped being clustered, and 15 s is the largest window possible given a 4 min long recording—we required 16 windows to ensure a stable estimate of clustering at the largest timescale. We chose 4 min as the minimal recording length to acquire the longest timescale possible while ensuring the availability of recordings (e.g. many popular song recordings are less than 5 min long).

The first step of AF analysis was to divide the downsampled signal into K segments, each one 4 min long, where K is number of adjacent segments needed to cover a given recording, including one aligned with the end of the recording and overlapping with the previous segment. For each segment, the envelope was computed using the Hilbert method (figure 1), although results were virtually identical (see electronic supplementary material) using half-wave rectification [21]. The latter is simpler to compute, and the two methods may not always yield the same results when signal bandwidths vary [22], but they yielded equivalent AF functions for peak events. Envelope peaks were identified in two steps. The first was to filter only peaks in the envelope that were maximal within ± 5 ms, and set all remaining points in the envelope to zero. We then zeroed out any retained peaks less than H in amplitude, and all remaining peaks were set to one. The threshold H was set such that one peak event was identified for every 200 samples in each recording, on average.

The ± 5 ms peak threshold served as a low pass filter because it set the maximal peak rate at 200 peaks per second. It served to bypass spurious peaks due to noise, while providing enough events to estimate clustering at the shortest timescale of 30 ms (two adjacent 15 ms windows). AF analysis requires a maximum event rate that is substantially higher than the shortest timescale to allow for sufficient variability in event counts per window. Preliminary tests indicated that a maximal rate of 200 Hz supported reliable AF variance estimates across all timescales measured—additional low-pass filtering would result in less accurate and reliable estimates of clustering due to low event counts. Events near the maximal rate of 200 Hz cannot be individually distinguished and related to perception or production, but they need not be, because clusters of events are the relevant units of analysis. We posit that sensory and motor systems may integrate over peak events to compute quantities like amounts of signal activity that are related to event clusters. But perception and production aside, our primary aim is to measure hierarchical temporal structure in the signals themselves.

The amplitude threshold H served to filter out noise by removing minor peaks from the analysis. It was set to create event series as sparse as possible, but with sufficient numbers of events to yield robust statistics. Finally, H also served to normalize recording levels in effect, by setting the amplitude threshold to yield roughly the same number of events per unit time across all recordings. If we set H to a specific decibel level

across all recordings, then the number of peaks identified would be influenced by arbitrary recording conditions among other factors. Nevertheless, analyses reported in the electronic supplementary material show that the results are robust to moderate changes in threshold settings [18,19].

Peak event series were submitted to AF analysis (figure 1), which originally was developed to quantify clustering in event times beyond chance, and has been used to measure clustering in neuronal spike trains [23]. AF analysis computes the Haar wavelet variance at a given timescale T by tiling an event series with windows of size T, and counting the number of events N within each window (figure 1). Differences in counts between adjacent windows are squared and averaged, and divided by twice the mean count, yielding an estimate of AF variance $A(T)$. Dividing variance by the mean is a type of coefficient of variation, similar to detrended fluctuation analysis [24], and it serves to normalize for the average event rate at each give timescale. This normalization helps to reduce the influence of factors like overall pitch and tempo on the pattern of hierarchical clustering in AF functions.

$A(T)$ was computed over three orders of magnitude, from $T \sim 30$ ms to $T \sim 30$ s, with 11 values of T in between, logarithmically spaced to compute the orthonormal basis. The shortest timescale is near the auditory flutter fusion threshold of approximately 30 Hz [25,26], and shifting the window sizes by small amounts has no appreciable effect on results. Therefore, the smallest clusters analysed were big enough to be individually perceived, and at least for speech, individually controlled in articulatory production because phonetic features like plosivity are on this timescale. If there is no clustering of events beyond chance, then events are Poisson distributed and $A(T) \sim 1$ for all T. Similarly, if events are periodically distributed, then $A(T)$ will approach zero for timescales larger than the period. If events are clustered across timescales, then $A(T) > 1$ and increases with T. If clusters are nested self-similarly across timescales, then $A(T)$ scales up as a power law, $A(T) \sim T^{\alpha}$, where $\alpha > 0$. If clustering generally scales as a power law but drops off beyond some timescale, as in short-range correlation, then $A(T)$ will flatten out beyond that timescale.

AF analysis is akin to spectral analysis for a point process, and applying it the Hilbert envelope is akin to computing the modulation spectrum, which is based on the autocorrelation of the amplitude envelope. The modulation spectrum has been examined for the purpose of automatic speech recognition [27], uncovering the neural bases of speech processing [28], and music classification [29]. The modulation spectrum can be computed as the spectrum of the Hilbert envelope, which makes it similar to AF analysis of peaks in the Hilbert envelope—both analyses reflect the amount of energy at different frequencies of the amplitude envelope. However, the AF function for a given signal is not the same as its modulation spectrum over a given range of timescales. The reason is that peak extracting is a kind of signal reduction relative to the modulation spectrum. A large amount of detailed information about the amplitude envelope is removed in the peak event series, and peaks are equalized to one. Equalization serves to normalize amplitudes, as well as reduce the signal.

The modulation spectrum has been used to compare different categories of sounds in at least two previous studies. Singh & Theunissen [30] compared speech, Zebra Finch songs and natural sounds such as those of fires and streams. The authors found differences among these categories in joint spectral and temporal statistics. They also found that, with respect to the average modulation spectra, all three categories similarly followed a power law analogous to the AF power law described above. Ding and colleagues [21] also computed average modulation spectra, but for speech versus music. Spectra for speech showed a consistent peak around 5 Hz, whereas spectra for music showed a consistent peak around 2 Hz. Numerous differences between the studies might explain the difference in results.

We formulated our method of analysis as an alternative to the modulation spectra designed to extract and normalize hierarchical temporal structure for a broad range of complex acoustic signals. As a point of comparison, we conducted the same analyses for the modulation spectrum as we did for AF analysis, over roughly the same range of timescales. Specifically, the fast Fourier transform (FFT) was computed for each of the 4 min windows of the Hilbert envelope used in AF analyses. The resulting spectral power (squared amplitude) estimates were averaged within logarithmically sized frequency bins, akin to computing AF variance at timescales spaced by powers of two. The resulting modulation spectra are plotted in the electronic supplementary material, and they are broadly similar to the AF functions reported below. However, the modulation spectra are more irregular and idiosyncratic than AF functions because of they retain more detail from the original signal. Here we focus on AF functions because their regularity facilitates quantification, interpretation and comparison.

3. Results

3.1. Allan factor

All types of vocalizations and musical performances yielded clustered event series. This clustering was often plainly visible, as shown in figure 2 for four representative segments of recordings, one from each of the four main categories. AF analysis yielded a clustering function $A(T)$ for each recording, and the means of these functions for each subcategory are plotted in the four panels of figure 3 on logarithmic coordinates. The most general result is an overall trend for $A(T)$ to increase with timescale for all complex signals, which indicates a general property of nested clustering. Beyond this overall trend, categories differed in how $A(T)$ increased with timescale, and these differences shed light on the behavioural and social processes that underlie speech, song and music. The statistical reliability of differences is visible in the 95% confidence intervals shown in figure 3 [31].

Most broadly, all human-generated signals showed reliable increases in $A(T)$ at the longer timescales of seconds to tens of seconds. By contrast, synthesized speech and most animal vocalizations showed some decrease in $A(T)$ in the longer timescales (top left figure 3), which indicates an upper limit to nested clustering. This difference shows that human speech and music are more hierarchically structured than synthesized speech and most animal vocalizations, particularly at timescales of seconds to tens of seconds. Regarding synthesized speech, the lack of structure appears to stem from a lack of richness in prosodic variation because synthesizers generate prosody using impoverished correlates of meaning and intent. Consistent with this interpretation, Falk & Kello [15] recently showed that hierarchical temporal structure is enhanced in infant-directed speech compared with adult-directed speech, the former being associated with rich, exaggerated prosodic variation [32]. Together these results provide clear evidence for prosody as a factor in creating the nested clustering of peak events in speech.

Our analyses do not shed light on whether the decrease in $A(T)$ for some animal vocalizations relates to prosody, although some bird and whale songs have been considered in terms of their prosodic structure [33]. It is particularly

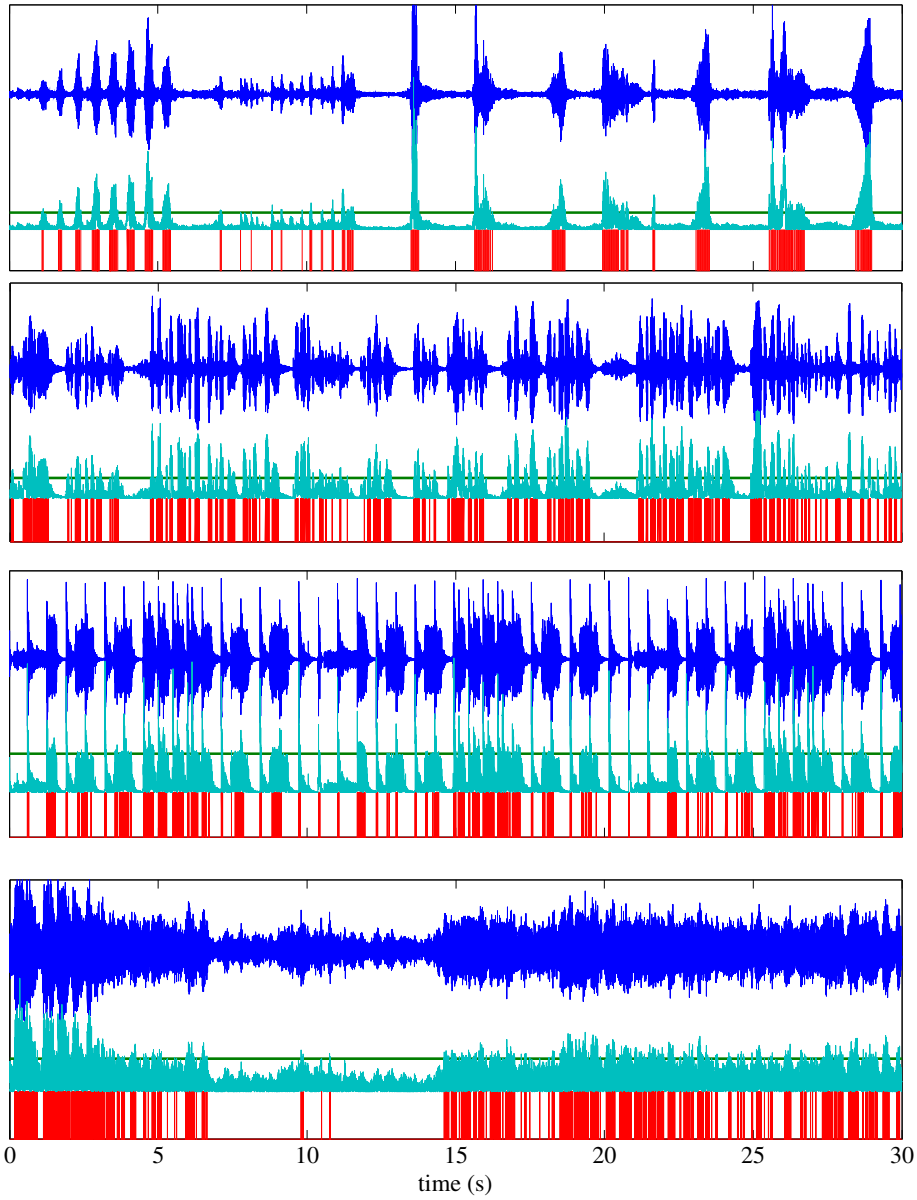

Figure 2. Four waveform segments, with corresponding Hilbert envelopes and peak event series, for one example recording from each of the four main signal types: Humpback whale song (top), TED talk (upper middle), rock music (lower middle; 'Back in Black' by ACDC) and symphony (bottom; Brahms symphony number 4, first movement).

interesting to note that the songs of humpback whales and hermit thrushes were highly similar in terms of their hierarchical temporal structure (middle right figure 3). This similarity is striking given such large differences in the environments, vocal apparatus and anatomy of these species. Further investigation is needed into prosody and behavioural factors that might explain such an unexpected similarity.

One such factor is the degree of social interaction reflected in recorded bird and whale vocalizations. The songs of humpbacks and hermit thrushes are solitary, as are nightingales. Only solitary male humpbacks sing long complex songs, and they mostly sing one song which changes over long periods of time. As for bird song recordings, they were collected from individual birds, with minimal interactions or singing from other birds (sometimes faintly in the background, but below the peak amplitude threshold). By contrast, killer whale vocalizations reflect their social interactions [34], and their recordings exhibited nested clustering that was surprisingly similar to human speech—conversational speech in particular. We hypothesize that this difference was observed because killer whale recordings reflected their social interactions whereas other animal recordings were solitary. This result suggests that hierarchical temporal structure is enhanced by social interaction as well as prosodic variation.

The hypothesized effect of social interaction is further supported by examination of AF functions for different types of human vocalizations. Just as nested clustering dropped off at longer timescales for solitary animal vocalizations, it also tapered off for monologue TED talks compared with conversational interviews, regardless of the language spoken (bottom left figure 3). We can explain this effect of social interaction by considering the sound of only one side of a conversation, like overhearing someone talking on the phone. The sound will have varying periods of silence during which the other person is speaking. If the temporal patterning of these periods is non-random and non-periodic, then it will create and enhance nested clustering in speech energy. The interview recordings primarily captured the interviewee's voice alone, which means that periods of

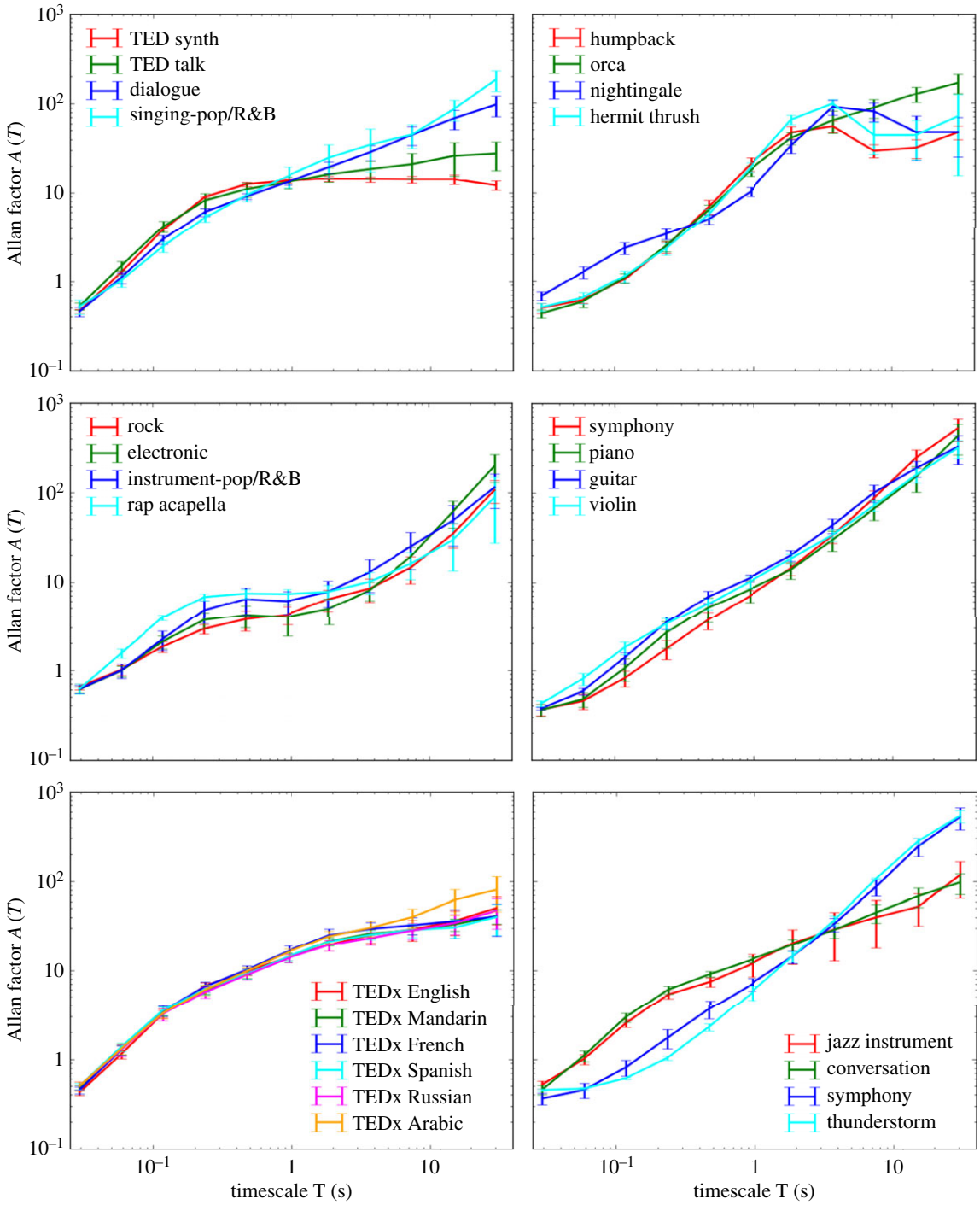

Figure 3. Mean $A(T)$ functions for all subcategories of signals analysed, with error bars representing 95% confidence intervals around mean AF variances for the sample of recordings within each subcategory, at each timescale. Power law AF functions would appear as positively sloped straight lines.

relative silence reflected turn-taking and other dynamics of coordination between interlocutors. Similarly, pop/R&B vocal recordings only captured the singer's side of interactions and coordination with other musicians and instruments, and their AF functions were strikingly like those of conversations (top left figure 3). Taken together, the evidence indicates that interaction dynamics enhance hierarchical temporal structure on timescales of seconds to tens of seconds.

Another factor that appeared to increase $A(T)$ at the longer timescales is musical structure. In general, AF functions were steeper in the longer timescales for musical recordings compared with speech or animal vocalizations (top right figure 3). This difference suggests that the hierarchical structure typical of musical compositions—roughly from phrases to melodies to periods—imparts even more nested clustering in peak events compared with the effects of prosody and social interaction. It is worth noting that, on top of the structure of a musical composition, musical recordings also capture interactions among musicians, akin to social interactions. These interactions do not explain steeper AF functions for music compared with conversational interactions.

The effects of musical structure and performance are further supported by the finding that nested clustering was most prominent across timescales for classical music recordings (middle left figure 3). This finding may stem from the fact that classical music possesses hierarchical structure in the melodic and metrical complexity conveyed by different instruments at different timescales, as compared with popular

music. In fact, classical music exhibited a power law increase in $A(T)$ as a function of T (i.e. linear in logarithmic coordinates), which indicates balanced nesting of events across timescales (for similar results based on the timing of notes, see [35]). The proportion of variance accounted for, R^2, was computed for each linear regression line fit to each classical music AF function in log–log coordinates. The mean R^2 was 98.4% for three orders of magnitude in timescale, which is strong evidence for power law AF functions.

In contrast with a power law, the temporal structure of popular music is often dominated by its rhythmic beat structure. As a result, their AF functions plateaued around the typical 1–2 beats per second. Note that AF analysis is not designed to pinpoint beat periods or other temporal rhythms at precise frequencies. Steady beats result in evenly timed, alternating periods of relatively few versus many events, which reduce $A(T)$ variance at timescales near the period of alternation. Jazz provided an interesting contrast—AF functions for jazz instrumentals closely followed that of conversations, rather than popular or classical music (bottom right figure 3). This exception is consistent with the adage that jazz is like a conversation [36], and may partly emerge from similar complexities in the temporal structure of jazz and speech, relative to popular music.

Finally, we applied our method to thunderstorms as a non-biological reference sound predicted to exhibit nested clustering in peak events. This prediction is based on the observation that thunder often occurs in long bouts, with shorter bursts of sound nested within bouts, and shorter crackles and booms further nested within bursts [37]. As predicted, AF analyses yielded nested clustering for three example thunderstorm recordings (bottom right figure 3). Intriguingly, their $A(T)$ functions followed a power law function very similar to that of classical music, and in particular, symphonies. Thus, it appears that nested clustering of events can be balanced across timescales as a product of human composition and performance, and also as a product of natural interactions across timescales.

3.2. Classification

Results thus far reveal a taxonomy of complex acoustic signals based on their patterns of clustering. This taxonomy is defined by three main dimensions of variation in AF functions: The degree of overall nesting as expressed by the slope in $A(T)$ as a function of T, the degree of floor or ceiling effects as expressed by the amount and direction of inflection in $A(T)$, and the degree of reduced nesting in middle timescales as expressed by the amount of middle flattening in $A(T)$. These three dimensions of variation can be neatly captured by fitting a third-order polynomial to each $A(T)$ function in logarithmic coordinates. The linear, quadratic and cubic coefficients become measures of the slope, inflection and flattening in $A(T)$. Analyses confirmed that polynomial fits captured the data, explaining about 99% of the variance for each $A(T)$ function. However, further analyses showed that three degrees of freedom were actually more than necessary. Correlations among coefficients suggested lower dimensionality in the data, and principal components analysis indicated that the first two components were sufficient to capture 99% of the variance. These components are largely reflected in the linear and quadratic coefficients, so for the sake of interpretability, individual recordings were plotted in figure 4 as points in two-dimensional coefficient space.

The scatter plot shows that the four main categories of complex acoustic signals were well-differentiated in terms of linear and quadratic features of their $A(T)$ functions. Animal vocalizations had steep and inflected $A(T)$ functions compared with human vocalizations, and the same was true for classical music compared with popular music. The two vocal categories were also separated from the two musical categories, in that animal and human vocalizations had lower quadratic coefficients than classical and popular music, respectively. To quantify separability, we trained and tested two linear support vector machines using fivefold cross-validation to classify individual $A(T)$ functions for two different levels of categorization: a superordinate level of categorization using the four main categories (40 recordings per category), and a subordinate level using all 16 subcategories shown in figure 4 (10 recordings per subcategory).

Classification performance showed that hierarchical temporal structure, as expressed in the shapes of AF functions, is highly diagnostic of their natural category. Classifiers were 92% correct on average (25% chance baseline) for the four superordinate categories, and 63% correct (6.25% chance) for the 16 subordinate categories. The different categories of recordings may appear easily discriminable in figure 2, but classification also requires generalization over variations within categories that are not shown. Classification results demonstrate that AF functions serve both to relate and distinguish different types of vocalizations and musical performances.

The added value of AF analysis is further bolstered by comparison with classification performance using two basic spectral measures, its mean and width (full width at half maximum), instead of temporal measures like AF linear and quadratic coefficients. Classifiers were only 55% correct on average, with animal vocalizations and classical music being somewhat more discriminable. Spectral mean and width are simple measures and there may be better spectral measures for classification. However, with no theoretical guide in hand, we leave it to future research to explore other measures. The present results allow us to conclude that nested clustering of peak events reflects a natural taxonomy of complex acoustic signals based on their hierarchical temporal structures.

4. Discussion

Hierarchical temporal structure is a fundamental property of speech, language, music and complex animal vocalization [38], but it has been challenging to measure and relate structures across different vocal and musical signals. In the present study, we formulated a method for measuring temporal hierarchies in complex acoustic signals from different domains. We found that patterns of event clustering across timescales could be reduced to points in an AF feature space, which allowed us to relate signals both within and across domains. We defined the space in terms of deviations from, and resemblance to, a power law relating AF variance and timescale. AF analysis is closely related to spectral analysis, and researchers have found similar power laws in spectral analyses of speech and music [35,39]. A power law in AF variance is indicative of nested event clustering that is self-similar across the

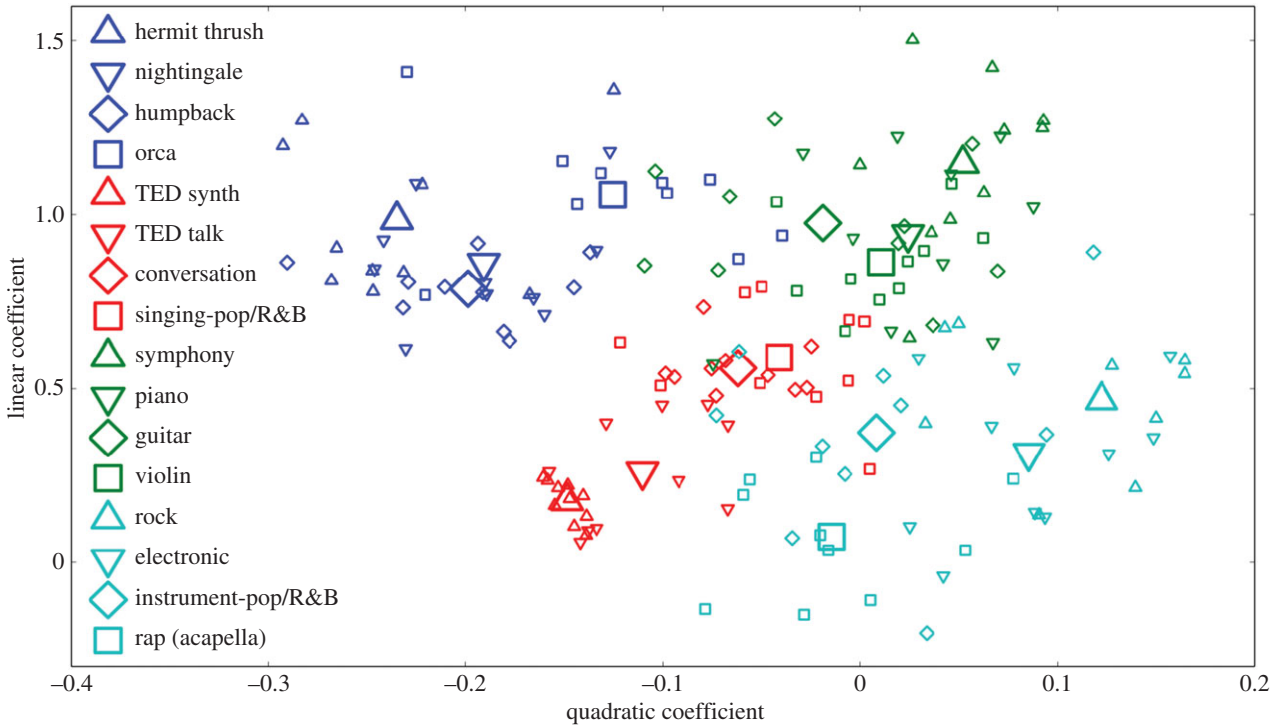

Figure 4. Scatter plot with each point representing the polynomial fit to the $A(T)$ function for a given individual recording. The linear and quadratic coefficients were taken from third-order polynomial fits to individual AF functions in log–log coordinates. The four main categories are represented by colour (blue = animal vocalization, red = human vocalization, green = classical music and cyan = popular music). Large symbols are placed at the centroid of each subcategory. Power laws have quadratic coefficients of zero and linear coefficients greater than zero.

range of timescales measured [40], and deviations from a power law indicate the relative enhancement or suppression of clustering at particular timescales.

Based on over 200 sound recordings, we found that enhancement and suppression of clustering revealed a taxonomy of sounds grouped by the behavioural and social processes involved in generating them. The superordinate categories reflected natural domains—human versus animal vocalizations, and popular versus classical music. The more specific subcategories reflected properties of vocal and musical production and interaction. Organization of the taxonomy yielded several useful and informative results. In terms of usefulness, methods for automatic classification of musical genres, and for distinguishing music from speech, have been developed for various music and speech applications [41,42]. The observed taxonomy illustrates how hierarchical temporal structure may be used to aid in numerous possible classifications among the categories and subcategories of sounds analysed [43]. Our method of AF analysis is simple and automatic, with the caveat that minutes-long recordings are needed to acquire data for the longer timescales.

Peak events in the amplitude envelope were chosen as acoustic features whose clustering may reflect the hierarchical temporal structures of processes that generate music and complex vocalizations. Peak events were not chosen as perceptual features, although the amplitude envelope is relevant to auditory perception [44], and the clustering of peak events may correlate with salient auditory features like temporal contrasts [45]. But in terms of generative processes, analyses identified three main factors that drove up the degree of nested clustering: prosodic variation, social interaction and musical structure. These all appear to be general sources of hierarchical temporal structure in the dynamics of vocalization and musical performance. Evidence for their generality is found in how AF functions showed commonalities among different types of speech and music, as opposed to the differences useful for classification.

First, prosodic variation had the same kind of effect on nested clustering regardless of whether it stemmed from infant-directed speech or synthesized speech. It also appeared to have a consistent effect on TED talks regardless of language spoken. This uniformity across languages is interesting because different languages are hypothesized to conform to different rhythmic patterns. Spanish, French and Mandarin are categorized as *syllable-timed* languages, whereas English, Arabic and Russian are categorized as *stress-timed* languages [46]—intervals between either syllables or stressed syllables are hypothesized to be roughly constant in duration. As noted earlier, constancy in timing means a lack of nested clustering, so one might expect stress timing and syllable timing to create plateaus in AF functions, like those associated with the beat structure of popular music. However, no such plateaus were observed in speech recordings, indicating that hypothesized effects of syllable/stress timing were either absent or not detectable with AF analysis. Instead, AF functions appeared to reflect a style of presentation common to TED talks, in that speakers of any language may use a common register in performing a rehearsed lecture designed to engage and inform. More work is needed to investigate this interpretation, but we can conclude that AF functions captured hierarchical temporal structure common to human speech, regardless of the inventories of phonemes, syllables, words, and sentences being spoken, and regardless of the cultural context.

Second, we found that interaction dynamics had the same effect on nested clustering regardless of whether interactions were among whales, speakers or musicians. Specifically, nested clustering was greater for conversational interviews

versus monologue TED talks, for musical interactions versus solitary vocal tracks, and for killer whales versus humpback whales. The unifying effect of interaction dynamics was even more stark and direct in the similarities between jazz instrumentals, spoken conversations and killer whale interactions. These results suggest further investigations into factors like turn-taking and coordination dynamics [47] that may have general effects on the dynamics of social interactions. Such investigations may use AF analysis as well as other complementary techniques like wavelet coherence analysis [48] and recurrence quantification analysis [49] that focus on more specific timing relations.

Third, similar musical compositions had similar AF functions regardless of singer, musical instrument, musical surface features such as pitch that do not directly relate to the hierarchical temporal structure of a performance. AF functions for songs from popular music reflected their similar structure characterized by a prominent beat, which suppressed nested clustering near the period of the beat. Songs from classical music reflected common structural complexity in terms of self-similar nested clustering (e.g. structure of grouping) across the range of measured timescales. The degree of self-similar clustering, as measured by the slope of AF functions, was greatest in symphonies, and intriguingly similar to the sounds of thunderstorms. On the one hand, this last similarity may seem most puzzling because symphonies and thunderstorms are so unalike in so many ways. On the other hand, the power laws observed in these two signals are akin to another power law known as $1/f$ noise that is common throughout nature [50]. It is informative to consider some hypothesized explanations of $1/f$ noise that might shed light on self-similar clustering in classical music and thunderstorms.

One hypothesized explanation for $1/f$ noise is that power law dynamics reflect a balance between predictability and surprise in temporal structure that is aesthetically pleasing [51]. In this case, the power law in thunderstorms [52] would reflect an aspect of nature that humans perceive as aesthetically pleasing, and reproduce in various artistic forms [53]. A related hypothesis is that power law structure and dynamics are ubiquitous in nature, and human systems have evolved to adapt to these power laws by mirroring them in various respects [54]. A third hypothesis is that power laws reflect the tendency of complex systems to be poised near phase transitions between ordered and disordered states [55]. Several studies have reported evidence that the connectivity of neural systems is hierarchically power law structured [56–58], and temporal interactions across the hierarchy may give rise to power laws in neural and behavioural activity [59–62], including music and complex vocalizations.

The present study was not designed to address hypotheses about the origins of power laws in hierarchical temporal structure, and none of them seem to readily explain why only classical music and thunderstorms were found to exhibit these power laws in their AF functions. Further investigation is needed, but we note that both are characterized by multiple timescales of interacting processes, stemming from either the structure of grouping and performance in classical music, or natural processes in thunderstorms. Some of these processes are neural and behavioural, as noted above, and they may also include social processes as well. Power laws can emerge as a general property of interactions across timescales, irrespective of the components and processes supporting these interactions [63].

In conclusion, we have developed a simple method for analysing nested clustering in a wide variety of complex acoustic signals. Results go beyond previous studies in their ability to clearly delineate and relate the hierarchical temporal structures underlying human speech and music, and complex animal vocalizations as well. Further investigation is needed to examine whether AF functions may reflect other aspects of behavioural and social processes not tested herein that underlie complex vocalizations and musical performances. New insights may be uncovered about the relationships between human speech and music, and other complex animal vocalizations.

Ethics. There was no human or animal experimentation, and no known risks to the researchers.

Data accessibility. Code and the list of recordings analysed are available at http://cogmech.ucmerced.edu/downloads. Sound files are freely available at YouTube, the Macaulay library (https://www.macaulaylibrary.org), TED talks (https://www.ted.com/talks) or the Buckeye speech corpus (http://buckeyecorpus.osu.edu). Researchers may also obtain recordings by contacting the first author at ckello@ucmerced.edu.

Authors' contributions. C.T.K. conceived the study and collaborated with S.D.B. on initial analyses and stimuli choices. B.M. also gathered stimuli and conducted some of the analyses, along with C.T.K. R.B. consulted on some analyses, and all four authors were involved in writing the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This research was funded in part by a Google Faculty Award to C.T.K., and a junior grant from the Institut Universitaire de France (IUF) to S.D.B.

Acknowledgements. The authors would like to acknowledge members of the Euromov Institute (http://euromov.eu/home) and the Cognitive and Information Sciences group (http://cogsci.ucmerced.edu) for feedback on various aspects of the work.

References

1. Hebets EA, Papaj DR. 2005 Complex signal function: developing a framework of testable hypotheses. *Behav. Ecol. Sociobiol.* **57**, 197–214. (doi:10.1007/s00265-004-0865-7)
2. Weiss M, Hultsch H, Adam I, Scharff C, Kipper S. 2014 The use of network analysis to study complex animal communication systems: a study on nightingale song. *Proc. R. Soc. B* **281**, 20140460. (doi:10.1098/rspb.2014.0460)
3. Kershenbaum A *et al.* 2016 Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biol. Rev.* **91**, 13–52. (doi:10.1111/brv.12160)
4. Baniya BK, Ghimire D, Lee J (eds). 2015 Automatic music genre classification using timbral texture and rhythmic content features. In *2015 17th International Conference on Advanced Communication Technology (ICACT), Seoul, South Korea, 1–3 July 2015*, pp. 434–443. Piscataway, NT: IEEE.
5. Shriberg E *et al.* 1998 Can prosody aid the automatic classification of dialog acts in conversational speech? *Lang. Speech* **41**, 443–492. (doi:10.1177/002383099804100410)
6. Harma A (ed). 2003 Automatic identification of bird species based on sinusoidal modeling of syllables. In *Acoustics, Speech, and Signal Processing, 2003*

Proceedings (ICASSP '03) 2003 IEEE International Conference on; Hong Kong, 6–10 April 2003, pp. 545–548. Piscataway, NJ: IEEE.

7. Rohrmeier M, Zuidema W, Wiggins GA, Scharff C. 2015 Principles of structure building in music, language and animal song. *Phil. Trans. R. Soc. B* **370**, 20140097. (doi:10.1098/rstb.2014.0097)
8. Koelsch S, Rohrmeier M, Torrecuso R, Jentschke S. 2013 Processing of hierarchical syntactic structure in music. *Proc. Natl Acad. Sci. USA* **110**, 15 443–15 448. (doi:10.1073/pnas.1300272110)
9. Martin JG. 1972 Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychol. Rev.* **79**, 487–509. (doi:10.1037/h0033467)
10. Cholewiak DM, Sousa-Lima RS, Cerchio S. 2013 Humpback whale song hierarchical structure: historical context and discussion of current classification issues. *Mar. Mamm. Sci.* **29**, E312–E332. (doi:10.1111/mms.12005)
11. Yu AC, Margoliash D. 1996 Temporal hierarchical control of singing in birds. *Science* **273**, 1871–1875. (doi:10.1126/science.273.5283.1871)
12. Goldsmith JA. 1990 *Autosegmental and metrical phonology*. New York, NY: Basil Blackwell.
13. Lerdahl F, Jackendoff R. 1983 An overview of hierarchical structure in music. *Music Percept. Interdiscip. J.* **1**, 229–252. (doi:10.2307/40285257)
14. Leong V, Goswami U. 2015 Acoustic-emergent phonology in the amplitude envelope of child-directed speech. *PLoS ONE* **10**, e0144411. (doi:10.1371/journal.pone.0144411)
15. Falk S, Kello CT. 2017 Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition* **163**, 80–86. (doi:10.1016/j.cognition.2017.02.017)
16. Allan DW. 1966 Statistics of atomic frequency standards. *Proc. IEEE* **54**, 221–230. (doi:10.1109/PROC.1966.4634)
17. Abney DH, Paxton A, Dale R, Kello CT. 2014 Complexity matching in dyadic conversation. *J. Exp. Psychol. Gen.* **143**, 2304–2315. (doi:10.1037/xge0000021)
18. Luque J, Luque B, Lacasa L. 2015 Scaling and universality in the human voice. *J. R. Soc. Interface* **12**, 20141344. (doi:10.1098/rsif.2014.1344)
19. Torre IG, Luque B, Lacasa L, Luque J, Hernández-Fernández A. 2017 Emergence of linguistic laws in human voice. *Sci. Rep.* **7**, 43862. (doi:10.1038/srep43862)
20. Pitt MA, Johnson K, Hume E, Kiesling S, Raymond W. 2005 The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Commun.* **45**, 89–95. (doi:10.1016/j.specom.2004.09.001)
21. Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. In press. Temporal modulations in speech and music. *Neurosci. Biobehav. Rev*.
22. Flanagan JL. 1980 Parametric coding of speech spectra. *J. Acoust. Soc. Am.* **68**, 412–419. (doi:10.1121/1.384752)
23. Teich MC, Lowen SB. 1994 Fractal patterns in auditory nerve-spike trains. *IEEE Eng. Med. Biol. Mag.* **13**, 197–202. (doi:10.1109/51.281678)
24. Kantelhardt JW, Koscielny-Bunde E, Rego HHA, Havlin S, Bunde A. 2001 Detecting long-range correlations with detrended fluctuation analysis. *Physica A* **295**, 441–454. (doi:10.1016/S0378-4371(01)00144-3)
25. Harbert F, Young IM, Wenner CH. 1968 Auditory flutter fusion and envelope of signal. *J. Acoust. Soc. Am.* **44**, 803–806. (doi:10.1121/1.1911177)
26. Lotze M, Wittmann M, von Steinbüchel N, Pöppel E, Roenneberg T. 1999 Daily rhythm of temporal resolution in the auditory system. *Cortex* **35**, 89–100. (doi:10.1016/S0010-9452(08)70787-1)
27. Kanedera N, Arai T, Hermansky H, Pavel M. 1999 On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Commun.* **28**, 43–55. (doi:10.1016/S0167-6393(99)00002-3)
28. Giraud A-L, Poeppel D. 2012 Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517. (doi:10.1038/nn.3063)
29. Lim SC, Jang SJ, Lee SP, Kim MY (eds). 2011 Music genre/mood classification using a feature-based modulation spectrum. In *International Conference on Mobile IT Convergence, Gyeongsangbuk-do, South Korea, 26–28 September 2011*, pp. 133–136. Piscataway, NJ: IEEE.
30. Singh NC, Theunissen FE. 2003 Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* **114**, 3394–3411. (doi:10.1121/1.1624067)
31. Nakagawa S, Cuthill IC. 2007 Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* **82**, 591–605. (doi:10.1111/j.1469-185X.2007.00027.x)
32. Fernald A, Taeschner T, Dunn J, Papousek M, De Boysson-Bardies B, Fukui I. 1989 A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J. Child Lang.* **16**, 477–501. (doi:10.1017/S0305000900010679)
33. Yip MJ. 2006 The search for phonology in other species. *Trends Cogn. Sci.* **10**, 442–446. (doi:10.1016/j.tics.2006.08.001)
34. Janik VM. 2009 *Acoustic communication in delphinids. Advances in the study of behavior*, vol. 40, pp. 123–157. New York, NY: Academic Press.
35. Levitin DJ, Chordia P, Menon V. 2012 Musical rhythm spectra from Bach to Joplin obey a 1/f power law. *Proc. Natl Acad. Sci. USA* **109**, 3716–3720. (doi:10.1073/pnas.1113828109)
36. Sawyer RK. 2005 Music and conversation. In *musical communication*, vol. 45 (eds D Miell, R MacDonald, DJ Hargreaves), pp. 45–60. Oxford, UK: Oxford University Press.
37. West BJ, Shlesinger M. 1990 The noise in natural phenomena. *Am. Sci.* **78**, 40–45.
38. Patel AD. 2003 Language, music, syntax and the brain. *Nat. Neurosci.* **6**, 674–681. (doi:10.1038/nn1082)
39. Voss RF, Clarke J. 1975 '1/f' noise in music and speech. *Nature* **258**, 317–318. (doi:10.1038/258317a0)
40. Lowen SB, Teich MC. 2005 *Fractal-Based point processes*. New York, NY: John Wiley.
41. Saunders J (ed) 1996 Real-time discrimination of broadcast speech/music. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, 7–10 May 1996*, pp. 993–996. Piscataway, NJ: IEEE.
42. Scaringella N, Zoia G, Mlynek D. 2006 Automatic genre classification of music content: a survey. *IEEE Signal Process Mag.* **23**, 133–141. (doi:10.1109/MSP.2006.1598089)
43. Dalla Bella S, Peretz I. 2005 Differentiation of classical music requires little learning but rhythm. *Cognition* **96**, B65–B78. (doi:10.1016/j.cognition.2004.12.005)
44. Houtgast T. 1989 Frequency selectivity in amplitude-modulation detection. *J. Acoust. Soc. Am.* **85**, 1676–1680. (doi:10.1121/1.397956)
45. Kayser C, Petkov CI, Lippert M, Logothetis NK. 2005 Mechanisms for allocating auditory attention: an auditory saliency map. *Curr. Biol.* **15**, 1943–1947. (doi:10.1016/j.cub.2005.09.040)
46. Bertrán AP. 1999 Prosodic typology: on the dichotomy between stress-timed and syllable-timed languages. *Lang. Des. J. Theor. Exp. Linguist.* **2**, 103–130.
47. Oullier O, de Guzman GC, Jantzen KJ, Lagarde J, Scott Kelso JA. 2008 Social coordination dynamics: measuring human bonding. *Soc. Neurosci.* **3**, 178–192. (doi:10.1080/17470910701563392)
48. Lachaux J-P, Lutz A, Rudrauf D, Cosmelli D, Le Van Quyen M, Martinerie J, Varela F. 2002 Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiol. Clin.* **32**, 157–174. (doi:10.1016/S0987-7053(02)00301-5)
49. Webber CL, Zbilut JP. 2005 Recurrence quantification analysis of nonlinear dynamical systems. In *Tutorials in contemporary nonlinear methods for the behavioral sciences* (eds MA Riley, GC Van Orden), pp. 26–94. See http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.pdf.
50. Gardner M. 1978 Mathematical games: white and brown music, fractal curves and one-over-f fluctuations. *Sci. Am.* **238**, 16–32
51. Bernstein L. 1976 *The unanswered question: six talks at Harvard*. Cambridge, MA: Harvard University Press.
52. Féral L, Sauvageot H. 2002 Fractal identification of supercell storms. *Geophys. Res. Lett.* **29**, 1686. (doi:10.1029/2002GL015260)
53. Levitin DJ. 2006 This is your brain on music: the science of a human obsession. London, UK: Penguin.
54. Shepard RN. 1994 Perceptual-cognitive universals as reflections of the world. *Psychon. Bull. Rev.* **1**, 2–28. (doi:10.3758/BF03200759)
55. Bak P, Chen K. 1989 The physics of fractals. *Physica D* **38**, 5–12. (doi:10.1016/0167-2789(89)90166-8)

56. Bullmore E, Sporns O. 2009 Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198. (doi:10.1038/nrn2575)
57. Bressler SL, Menon V. 2010 Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.* **14**, 277–290. (doi:10.1016/j.tics.2010.04.004)
58. Eguíluz VM, Chialvo DR, Cecchi GA, Baliki M, Apkarian AV. 2005 Scale-free brain functional networks. *Phys. Rev. Lett.* **94**, 018102. (doi:10.1103/PhysRevLett.94.018102)
59. He BJ. 2014 Scale-free brain activity: past, present, and future. *Trends Cogn. Sci.* **18**, 480–487. (doi:10.1016/j.tics.2014.04.003)
60. Palva JM, Zhigalov A, Hirvonen J, Korhonen O, Linkenkaer-Hansen K, Palva S. 2013 Neuronal long-range temporal correlations and avalanche dynamics are correlated with behavioral scaling laws. *Proc. Natl Acad. Sci. USA* **110**, 3585–3590. (doi:10.1073/pnas.1216855110)
61. Kello CT, Beltz BC, Holden JG, Van Orden GC. 2007 The emergent coordination of cognitive function. *J. Exp. Psychol. Gen.* **136**, 551–568. (doi:10.1037/0096-3445.136.4.551)
62. Ihlen EA, Vereijken B. 2010 Interaction-dominant dynamics in human cognition: beyond 1/f(alpha) fluctuation. *J. Exp. Psychol. Gen.* **139**, 436–463. (doi:10.1037/a0019098)
63. Kello CT, Brown GDA, Ferrer-i-Cancho R, Holden JG, Linkenkaer-Hansen K, Rhodes T, Van Orden GC. 2010 Scaling laws in cognitive sciences. *Trends Cogn. Sci.* **14**, 223–232. (doi:10.1016/j.tics.2010.02.005)