An ancient evolutionary calculus for attention signaling retained in modern music

Gregory A. Babbitt[1] and Ernest P. Fokoue[2]

1. T.H. Gosnell School of Life Science, Rochester Institute of Technology
2. School of Mathematical Science, Rochester Institute of Technology

Abstract – The ultimate mechanism(s) underlying the evolution of human music have been the subject of both philosophical thought and scientific study. However, the proximate mechanism of fitness signaled through the quality of music has not. As music and dance functionally aim to hold an observer's attention, we investigate features in sound that might stimulate attentive behavior in an audience. We propose a calculus of attention to both movement and sound to ascertain its f(x)=control=position, f'(x)=energy=change in position, and f"(x)=surprise=change in direction (CES). We develop software to extract CES, (B) dynamically map audio trajectories within the space of CES, and (C) measure performer stability within CES (i.e. signaling performance quality). When musical sounds are compared to animal vocalizations, human speech, and non-biological sounds including noise, we confirm that music maintains the highest stability in the space of CES. Attention signaling also differs significantly between experts versus novice musicians, adult versus immature bird songs, and is constrained by performance interactions with live audiences as well. We conclude that quality signaling via this calculus of attention within behavioral display reflects physics of observing bodies in motion, probably originating from selective pressure of intra/interspecific conflicts formed at the Ediacaran-Cambrian boundary, and retained in music today.

Key words – evolution, music, attention, calculus, bird song, primate vocalization, machine learning

Lay Summary – Most animals (including humans) pay attention to specific audiovisual stimuli within their environment. Here we explore common features of these stimuli and propose a single mathematical framework (with software) for detecting and measuring the quality of performance regarding attention signaling in the context of sound. We show that musical sounds are exceptionally attractive to our attention and that attention signaling in music probably has ancient evolutionary roots that connect to when early life forms started to evolve swift directed movement and the sensory capacities to detect it.

Introduction

The origin, purpose, and nature of music have long been viewed through very different intellectual perspectives. Natural philosophers have repeatedly observed that music has roots in both cosmic and Earthly forms of nature (Kepler et al., 1997; Watkins, 2018), and like mathematics (Wigner, 1960), music seems to affect most people with respect to both their internal and external reality. More recently scientists across diverse fields have also explored the origin and function of music. Recent consensus across scientific disciplines might conclude that music contains features very close to and yet predating human language (Gustison et al., 2012; Gustison and Bergman, 2016; Peretz et al., 2015; Richman, 1987), is culturally diverse and yet rooted to some degree in genetics (Bignardi et al., 2025; Brown et al., 2014; Jacoby et al., 2024; Passmore et al., 2024), occurs in some animals outside humans (Baptista and Keister, 2005; De Gregorio et al., 2024; Fitch, 2015, 2006; Honing, 2018; Raimondi et al., 2023; Rothenberg et al., 2014; Snowdon and Teie, 2009) and even spontaneously within artificial intelligence (Kim et al., 2024), fosters social connection in the absence of shared language (Ravignani et al., 2014; Savage et al., 2021; Spitzer, 2021; Wolfe, 2002), and evokes stimulation and development across large parts of the human brain (Arjmand et al., 2017; Bedoya et al., 2021; Harding et al., 2025; Izbicki et al., 2020; Sacks, 2008; Trimble and Hesdorffer, 2017; Vuust et al., 2022). Anthropologists and biomusicologists have linked the evolution of music to Darwinian fitness (Adami, 2012; Huron, 2001; Patel, 2014) and the rise of bipedalism (Spitzer, 2021), the formation of prosocial behavior and larger societies (Pfattheicher et al., 2022; Savage et al., 2021), and the mental ability for cognitive abstraction via theory of the mind (Fitch, 2006; Spitzer, 2021). One perspective that has been less often applied to the subject of music is that of behavioral ecology. Here, we begin with basic questions about what fitness advantage is ultimately conferred to individuals engaging in musical behavior (Huron, 2001), and more specifically, what are the exact proximate mechanisms of signaling of individual quality (i.e fitness) involved in the display of musical behavior. While scientists have proposed some plausible ultimate biological and cultural drivers of musical behavior, to our knowledge, the proximate mechanisms by which quality is honestly signaled through musical performance has not been investigated. As the obvious aim of music and dance is to hold the attention of an audience (MacCallum et al., 2012), we propose that some basic investigation of acoustical features of music that arouse attentive behavior in the listener could prove especially insightful regarding the behavioral mechanism(s) that underlie the diverse perspectives and opinions regarding the evolutionary origin(s) of music.

The need for attentive behavior probably first arose during the transition from the Ediacaran to Cambrian period about 541 million years ago when animal life first evolved swift directed movements as well as the more complex brains and sensory organs required to detect

these movements. Associated with this E-C boundary are the rise of many new body forms, extensive burrowing behavior, and the evolution of protective shells covering soft body parts (Darroch et al., 2023; Nelson et al., 2022). Attentiveness to moving bodies can be described mathematically by a simple calculus where an observer must accurately gauge the position of another body in time and space (Figure 1A). This requires accurate reading of the control of position of the other organism (f(x)), its energy or velocity relative to the observer (f'(x)), and its change of direction through acceleration and deceleration relative to the observer (f''(x)). An incomplete ability of the observer to perfectly account for changes in direction would create an element of surprise and unpredictability on the part of the observer. We can also describe the quality of performance displays of an observed individual directly through their exhibition of motor control, kinetic or potential energy, and its cognitive reserve (i.e. ability to surprise). In the context of perhaps all behavioral displays of individual fitness across many animal species, including humans, we propose that this calculus of attention is always present to some degree, and that audiences of behavior displays of individual quality will be attentively engaged by any components of the display that also convey motor control, energy expenditure, and surprise (i.e. complexity) (Figure 1B). In human music and dance, we can readily observe elements of control, energy, and surprise (CES) and measure them in sound itself using a variety of audio processing methods (Figure 1C). We can mathematically define the CES of external sounds in a ternary space and even begin to map them onto a second ternary space of cognitive sensations they might evoke in the human mind (Figure 1D). By emphasizing the three components of CES individually or in pairs, we seem to also be able to describe the variety of cognitive effects we have from different musical genres, simple musical scales and the even art of the spoken word (Figure 2).

Here, we hypothesize that the more physically, energetically, and cognitively fit an individual musician or animal vocalist is, the better they can (A) control and stabilize their sound in the ternary space of CES and (B) move more intentionally within it (i.e. poor performance creates more randomness in the CES trajectory). We introduce an open-source, user-friendly software application named the POPSTAR PROJECT (https://gbabbitt.github.io/POPSTAR-signaling-in-music/) to dynamically map and analyze sound files (.wav or .mp3) in the space of CES. We introduce a stability indicator function for CES that measures performance with respect to the calculus of attention. We then compare (A) musical and non-musical sounds, (B) various animal vocalizations, (C) musical genres with different emphasis on solo performance, and (D) differences in musical performance ability and environment (e.g. expert vs. novice, live vs. studio). We find compelling evidence that this ancient attention signaling is representative of an

individual's physical, energetic, and cognitive ability and is still retained in most modern musical performance today.

Methods

The POPSTAR software begins with the audio file in .mp4 or .wav format and segments it into user defined windows spaced on either (A) every beat using python's librosa library beat tracker algorithm (Ellis, 2007) (if user checks the box labeled 'beats" on the GUI), or (B) every ¼ second if the user leaves 'beats' unchecked. We extract nine acoustic features of attention signaling from each audio segment, 3 each for control, energy, and surprise (i.e. CES). We define performance quality directly in terms of acoustic measurement of physical and cognitive indicators of individual quality (i.e. fitness) such as motor control, energy, and complexity. Note that this is quite different than mathematical descriptions of Darwinian fitness as heritable phenotypic changes in populations (e.g. the Price equation (Gardner, 2008)). While it is an obvious conjecture that direct indicators of physical and cognitive quality might translate to Darwinian fitness within populations, we do not measure or model this here.

Acoustic features of control (C)

Control of pitch or frequency

Pitch control (PC) is defined as

$$PC = \log \frac{1}{\sum_{i=1}^{N} abs\left(\frac{(O-E)^2}{E}\right)}$$

Where N = number of notes discovered with a f0 or fundamental frequency and O = f0 in Hz determined by librosa pYIN algorithm (de Cheveigné and Kawahara, 2002; Mauch and Dixon, 2014) and E = expected frequency of the nearest value on the western music scale at A440 Hz or else is determined by nearest cluster mean frequency determined via expectation-maximization (EM) clustering. If the user chooses music' or 'single' as sound-type on the GUI, then the Western music scale is used. If the user choses 'other' as sound-type on the GUI then the clustering approach is used.

Control of harmonics

Harmonics control (HC) is determined as the harmonic energy (HE) summed over N notes and L levels of harmonics above f0 detected by the librosa harmonic spectrum analyzer (Rafii, 2022; "Spectral Processing - DAFX - Wiley Online Library," n.d.).

$$HC = \log \sum_{i=1}^{N} \sum_{j=1}^{L} HE$$

Control of timing

Timing control (TC is determined as

$$TC = \log \frac{1}{\sum_{i=1}^{N} abs\left(\frac{(O-E)^2}{E}\right)}$$

Where O is observed beat interval and E is mean beat interval returned by librosa beat tracking software (Ellis, 2007)

Acoustic features of energy (E)

Tempo

Tempo (TP) is defined as

$$TP = B/t$$

where B = total beats in audio segment and t = time in minutes

Reverberation

Reverberation (RV) is determined as

$$RV = r_{L=1}$$

Where r is the Pearson correlation coefficient at lag (L) = 1 or the nearest peak to the center taken from the autocorrelation function (ACF) of the time domain audio signal. The ACF with time lag (k) is computed across the time domain (x) as

$$ACF = p(k) = \frac{cov[x_t, x_{t-k}]}{cov[x_t, x_t]}$$

Amplitude

The relative amplitude (AM) is determined as

$$AM = \log\left(\sum_{i=1}^{T} abs(A)\right)$$

Where A is the amplitude of the signal in time domain and T is the time length of the audio segment

Acoustic features of surprise (S)

Multi-scale entropy

The multi-scale entropy (MSE) is defined as the complexity index returned from the refined multi-scale entropy (rMSE) function for univariate data from python's EntropyHub library (Flood and Grimm, 2021) using an Entropy object with embedding dimension =4 and radius threshold = 1.25 to return a Complexity Index over 5 temporal scales using 3rd order Butterworth Filter with normalized corner frequency of 0.6 at each temporal scale.

Lempel-Ziv complexity

The Lempel-Ziv complexity (LZC) is defined as

$$LZC = \log N$$

Where N = number of distinct substrings in the binary representation of the audio file

Note variability index

Note variability is defined according to the method of (Sawant et al., 2022) for analyzing bird song complexity. The spectral cross-correlation is converted to a Note Variability Index (NVI) ranging between 0 and 1 where

$$NVI = \frac{\sum_{i=1}^{N_C} \sum_{j=1}^{N_C} \left(1 - C_{\Delta T_{I,j}}\right)}{N_C(N_C - 1)}$$

forming an N_C X N_C similarity matrix where N_C is the number of notes and $C_{\Delta T}$ is the association between two notes. Thus, a lack of cross-correlation across all pairs of notes or data columns in a spectrogram represents high variability in the sound file while a tone repeated constantly in the time domain has high cross-correlation but no complexity. It is important to understand that variability in the signal increases to the point of unpredictability, thus the variability eventually increases from a complex pattern to memoryless white noise (where the next value in a sequence cannot be predicted at all).

Min max (self) and Z score normalized average values of 9 features of CES

If the user chooses 'single' file under sound-type on the GUI, the CES feature matrix constructed listed above is min-max normalized by column where

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

If the user chooses 'music' under sound-type, the CES feature matrix is z-score normalizes using the average feature parameter values (μ,σ) for human music where

$$x' = \frac{x - \mu}{\sigma}$$

If the user chooses 'other' under sound-type, the CES feature matrix is z-score normalizes using the average feature values (μ,σ) for human speech.

Dynamic ternary plot representation of CES over time

Control, energy and surprise features for each audio segment in time (t) are averages of the normalized features from that audio segment

$$C_t = mean(PC, HC, TC)$$

$$E_t = mean(TP, RV, AM)$$

$$S_t = mean(MSE, LZI, TC)$$

and thus

$$(C_t, E_t, S_t), t = 1 + w, 2 + w, \dots T$$

where t is user defined as t = ¼ second OR beats tracked in the time of the music and w = analysis widow length determined by user (default is 8 seconds) and values of $(CES)_t$ are normalized to sum to 1

The single points in CES ternary space are thus

$$(C_t, E_t, S_t), t = 1,2, \dots T \xrightarrow{yields} P_t\ , t = 1,2, \dots T$$

A dynamic ternary plot is generated as well as a dynamic Chernoff face (Chernoff, 1973) and saved in movie file (.mp4) format (Figure 3). In the Chernoff face representation, Eye height = NVI, Eye width = LZC, Pupil area = MSE, Eyebrow slant level = TP, Mouth height = AM, Mouth width = RV, Ear height = PC, Ear width = TC, and Nose slant = HC. Examples of .mp4 (movies) containing dynamic Chernoff faces and dynamic ternary plots with sound overlay for a variety of bird song, human/animal vocalization, and human music can be found in examples_dynamicPlots.zip

Stability indicator (i.e. performance fitness indicator) for the tracking of CES over time

We develop a simple permutation-based indicator of the stability of the performer's trajectory within the space of CES (Figure 4). We hypothesize that high levels of performance quality and resulting musicality of sound (i.e. fitness of performance) should result in more stable and intentional movement within the ternary space of CES. To measure the stability of any given audio signal we create permutations of the observed CES

trajectory by shuffling time steps. Note this does not alter the values of points in the ternary space, only the order in which they occur, thus

$$(P_1, P_2, \dots P_T) \xrightarrow{yields} \left(P_{\sigma(1)}, P_{\sigma(2)}, \dots P_{\sigma(T)}\right)$$

Each observed time step in the ternary space

$$\delta_t = P_{t+1} - P_t$$

Is compared to each corresponding shuffled (i.e. random) time step

$$\delta_{\sigma(t)} = P_{\sigma(t+1)} - P_{\sigma(t)}$$

and we count instances in which the observed step is less than or greater than its corresponding shuffled step

$$\kappa(A) = \sum_{t=1}^{T} \mathbb{1}\left(\delta_t < \delta_{\sigma(t)}\right)$$

$$\kappa(B) = \sum_{t=1}^{T} \mathbb{1}\left(\delta_t \geq \delta_{\sigma(t)}\right)$$

Thus. an indicator of stability ranging from 0 to 100 percent in the CES trajectory becomes

$$\eta(A) = \left[\left(\frac{\kappa(A)}{\kappa(A) + \kappa(B)}\right) - 0.5\right] * 200$$

Nonparametric hypothesis testing on average C_t, E_t, S_t and stability = η(A) can be achieved using Kruskal-Wallis H test where

$$H = (N-1)\frac{\sum_{i-1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} n_i\left(r_{ij} - \bar{r}\right)^2}$$

Where *N* is total observations across all groups, *g* is number of groups, n_i is number of observations in group i, r_{ij} is rank of observation j from group i, $\bar{r}_i$ is average rank of all observations in group I and $\bar{r}$ is average of all r_{ij}

Multi-model inference can also be performed on the observed step distributions using Bayesian information criterion

$$BIC = -2\ln(L) + k\ln(n)$$

where L is likelihood of model given the data, k is the number of parameters in the model and n is the number of data points

Random forest classification

The 9 normalized acoustic features are used to train subsequent random forest models that were used to assess comparative classification of sound types derived from our protein sonification when compared within the context of human speech, music, and other natural and computer-generated sounds. All sound files used are listed in folder_list.txt and included in Supplemental Data. Note that some music files are omitted due to copyright, but the titles of the music tracks analyzed can be seen in the first column of featureData_allFolders.dat. Our random forests consisted of 500 decision trees using 80/20% training/testing data split. It is implemented with python scikit-learn RandomForestClassifier function which implements the ID3 with CART algorithm to create each decision tree. Splits on the trees are called using the Gini impurity or Gini index.

$$Gini\ impurity = \sum_{i-1}^{N} f_i(1 - f_i)$$

Where f_i is the frequency of label i at a node and N is the number of labels.

The importance for each feature (I) in the random forest is given by

$$I = \frac{\sum_{j=all\ trees} norm\ I_j}{T}$$

where the normalized feature importance is for a give tree is

$$norm\ I_{ij} = \frac{I_i}{\sum_{j=all\ features} I_j}$$

where the feature importance for a give tree is

$$I_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k=all\ nodes} ni_k}$$

where n_j the importance of each node j is

$$n_j = w_j G_j - w_{j(left)} G_{j(left)} - w_{j(right)} G_{j(right)}$$

where w is the number of samples reaching node j, G is the impurity value of node j and left(j) and right(j) indicate the child nodes of node j

Data retrieval and statistical comparisons

In our research, we use the POPSTAR PROJECT code, popstar.py available at (https://gbabbitt.github.io/POPSTAR-signaling-in-music/) to generate dynamic ternary maps and make statistical comparisons of CES and attention signaling (i.e. CES stability) across musical and non-musical sounds, musical genres, an artists discography, and selected examples of human, primate, bird, and whale vocalizations. Movie output for specific examples of the dynamic ternary plots and dynamic Chernoff faces can be observed at the website above. We make specific comparisons of attention signaling in expert vs. novice vocalists in both humans and Yellow Canary (owners often post the first songs of their pet canaries on YouTube after training them with videos of 'exceptional' adult birds singing). We compare attention signaling in piano concertos, where the ability of soloists is key, to ambient piano works of the Impressionist era, where the music emphasizes the environment of the listener rather than the skill of the soloist. We also compare live vs. studio performances of the popular artist Bjork Guomundsdottir, to see if direct feedback of the audience affects the resulting attention signaling of the artist.

In cases where audio was extracted from YouTube we avoided videos with excessive background noise and performed batch audio extraction and batch loudness normalization using our script (yt-dlp_batch.py) available in the POPSTAR Project Github repository. All ternary plot movie output (with sound) and statistical results are available at https://doi.org/10.5281/zenodo.17218769.

Results

In comparisons of stability of audio trajectory in the ternary space of CES (i.e. signaling of individual performance quality) across a large variety of musical and non-musical sounds (Figure 5), we find that music has the highest stability in CES (>60-70% stable) while samples of brown noise have near 0% stability (Figure 5A). Therefore, our indicator function for stability in CES is a valid measure of the human perception of musicality in sound. We find that musical genres have significantly different and higher CES trajectory stability than human speech ($H = 91.045$, $p < 0.001$) (Figure 5B) and that the career discography of the pop musician Bjork Guomundsdottir also exhibits significant variation in CES stability ($H = 85.492$, $p < 0.001$) (Figure 5C). Comparison of CES stability across human singing and a variety of primate, bird and whale vocalizations demonstrates significant differences ($H=29.320$, $p<0.001$) and also that some more musical sounding species such as

humpback whale and European Nightingale, can match or even surpass some humans in this measure of musicality (Figure 5D, Figure S1). We also observe that 3 species of sparrows show a gradient of CES stability, with Chipping Sparrows matching the musicality of human speech and the aptly named Song Sparrows matching the musicality of human singing, and White-Throated Sparrows falling in between (Figure 5D, Figure S1).

Comparisons of expert and novice human singers (Figure 6A-C) reveal that expert musicians exhibit borderline significantly more CES stability ($H=4.339$, $p=0.037$), more control ($H= 3.429$, $p=0.064$) and less surprise ($H=5.357$, $p=0.021$) than novice musicians (Figure 6A-B). A machine learning random forest classifier identifies expert vs. novice human singers with 99% accuracy using mostly the features of Lempel-Ziv (substring) complexity and note variability and amplitude modulation to a lesser extent (Figure 6C). In a similar analysis comparing adult Yellow Canary to immature birds singing their first songs (Figure 6D-F) we find no differences in overall levels of control, energy or surprise, but we do find that adults have significantly more musicality indicated by the stability in CES ($H=3.971$, $p=0.046$)(Figure 6E). Machine learning classification can distinguish adult from immature song with 93% accuracy using a wide variety of features of CES (Figure 6F).

As not all music places equal emphasis on solo performance (i.e. individual performance quality), we compared a small sample of the most difficult piano concertos for soloists to perform (Rachmaninov Nos. 2&3, Beethoven No. 5, Mozart Nos. 20&21, Brahms No. 1, Tchaikovsky No. 1, Grieg Am concerto) with a small sample of ambient piano pieces by the composers Satie (Gymnopedie 1-3, Gnossienne 1-6), and Debussy (Arabesque No. 1&2, Clair de lune, Reverie) (Figure 7). Here we find that while the levels of control energy and surprise are similar (Figure 7A), the stability in CES is significantly higher in the concertos designed to demonstrate soloist virtuosity than in the ambient pieces which emphasize music as a space which the listener occupies ($H= 5.357$, $p=0.021$) (Figure 7B). Random forest classification can discriminate concertos from ambient works on piano with 95% accuracy indicating that multiscale entropy and reverberation are the most important features for learning the difference here (Figure 7C).

Lastly, we analyze the career of the pop musician Bjork who is very interesting from the context of our research because of the large range of 'controlled' experimentation that she has recorded. Across a nearly 50-year career and 6 World tours, she is well known to be an excellent live vocalist performer and as well a highly experimental composer touching upon many unique genres (pop, punk, trip-hop, jazz, folk, electronica, ambient, and avant-garde). Her vocal exploration of the boundary between her own human speech and her more conventional music is also important for our study, particularly her fifth studio album Medulla, which attempts to produce fully modern music using only the human voice alone

(i.e. no instrumentation) (Figure S2). After each of her world tours she released a live album corresponding with the same tracks as her studio work prior to touring, thus providing a nicely controlled experiment where we can examine the effect of the feedback of live audiences on the different components of attention signaling in her performances (Figure 8). While we find no significant differences between live and studio performance regarding the stability of CES trajectories, we do see different levels of CES evident across 6 studio LPs recorded over 18 years (Figure 8A) that are absent in the corresponding live performance LPs (Figure 8B). This indicates more consistent combination of control, energy and surprise when facing a live audience. Combined comparison of live vs studio LP's (Figure 8C) finds significantly increased control ($H=11.918$, $p=0.001$) and decreased surprise ($H= 8.194$, $p=0.004$) again indicative that audience feedback greatly effects attention signaling in a musical performer. We also note that the differences between live vs. studio performance is not reliant on any single feature of CES (Figure 8D). Lastly, we compared Bjork's attention signaling across normal speech (i.e. TV interviews), her Medulla LP, and her Greatest Hits LP (Figure S2) and find significant differences in all levels of CES (control: $H=39.034$, $p<0.001$ | energy: $H=15.014$, $p=0.002$ | surprise: $H=26.961$, $p<0.001$) and stability of CES ($H=16.033$,$p=0.003$) with clear evidence from the Medulla LP indicating that the attention signaling present in the audio is not dependent upon electronic instrumentation, but can be the product of human vocalization alone.

Discussion

We have proposed a novel multivariate measure of vocal performance (i.e. individual quality) in sound defined directly at the phenotypic level through motor control, energy, and surprise/cognitive complexity. We propose that these three sets of acoustic features (i.e. control, energy and surprise or CES) combine to form a mathematical calculus of attention that individuals can manipulate to stimulate attentive behavior in an observer (or larger audience). We further propose that the quality of attention signaling can be indicated by the level of non-randomness in the trajectories of sounds in the ternary space of CES. We demonstrate that attention signaling defined in this way is common across many forms of human and animal vocalization and perhaps also associated with visual displays as well. We investigate auditory attention signaling in human and animal vocalizations compared to other non-biological sounds and demonstrate that it corresponds to our human perception of musicality in both sounds we culturally define as music, as well as natural sounds we perceive as more musical than others (e.g. song sparrows and nightingales). We demonstrate that expert vocalists, in both humans and some birds (Yellow Canary), have higher quality levels of attention signaling than novice individuals as evidenced by less randomness of song trajectory in the space of CES. This finding is supported by a recent study showing that daily vocal exercise is as important for birds as it is for human singers

(Adam et al., 2023) and therefore is likely to be an honest signal of fitness through performance as well. We also find attention signaling is stronger in musical performances that emphasize the ability of soloists over the perceptions or moods of the listener (i.e. piano concertos vs. ambient piano music). Lastly, we demonstrate that in the career of a famous popular performer, live audience feedback seems to constrain the diversity of her exploration of the space of CES when compared to work in the studio. This suggests there is something consistent to attention signaling that requires audience presence and helps to explain previous observations in neuroscience that show that people are often more stimulated by live vs. recorded music (Trost et al., 2024).

While some of our comparisons have only moderate sample sizes, the fact we are finding consistent functional differences among vocalizations and musical sounds regarding the stability of CES suggests that our mathematical definition of fitness in animal vocal performance and human music is useful and further investigations at the intersection of aesthetics and evolutionary biology could prove insightful. In future studies, we plan to investigate the utility of our software and methodology for predicting historic and current trends in the popularity of music (e.g. peak Billboard magazine chart positions, YouTube views, or Spotify downloads) and in training amateur musicians and/or designing new kinds of instrumental or electronic sounds that stimulate the attention of audiences in interesting ways. Musical sound stimulates the combined neurology of the human brain more than any other neurophysiological stimulus (Sacks, 2008). Within the context of music, our proposed biopsychological model hypothesizes the existence of an external representation in ternary space rooted in the calculus of our attention to musical CES. This might underpin our internal cognitive response to music in myriad combinations of physical, emotional, and intellectual feelings. Perhaps future work within the field of music and neurophysiology might be able to further confirm or reject the ternary architecture of our mathematical model.

It has been observed that human minds struggle to conceptualize statistical relationships that extend much beyond 3-dimensional space and have trouble tracking more than a handful of objects in real time (Ceja and Franconeri, 2023). Herman Chernoff invented facial representation of complex data because faces are one of the only multivariate structures that evolution has equipped us to rapidly and easily interpret (Chernoff, 1973). Recent work demonstrates that human imagination can only accurately predict the future moving path of one single object at a time (Balaban and Ullman, 2025). These psychological experiments on the limitations of human attention and imagination speak volumes about how our attentive behaviors were probably molded by our ancient evolutionary history. The Ediacaran-Cambrian transition 541 million years ago created selective pressures underlying a singular calculus of attention, thereby limiting the

dimensionality of our minds in ways that computer algorithms behind machine learning are not. Indeed when attention has been implemented in the transformer neural network architecture of modern generative artificial intelligence (GAI), it has vastly improved its capability to mimic and interact with or own neural architecture (Vaswani et al., 2023).

The concept of a calculus of attention has broad implications beyond popular music, especially in an emerging internet economy where people's attention is increasingly monetized for exploitation by GAI (Heitmayer, 2025; Vettehen and Schaap, 2023). We propose that investigating the underlying mathematical patterns and fundamental biology of attentive behavior can potentially help guard our society against GAI applications aimed at overstimulating the cognitive mechanisms that control our attention. Humans not only direct their attention towards popular music, but also popular athletes, public speakers, teachers, preachers, politicians, and newscasters, all of whom might also be exploiting behavior features designed by evolutionary biology to hold our attention. This can potentially affect us in ways that can prove either productive or unproductive to our lives depending upon the context with which we engage in internet outlets for modern media.

References

Adam, I., Riebel, K., Stål, P., Wood, N., Previs, M.J., Elemans, C.P.H., 2023. Daily vocal exercise is necessary for peak performance singing in a songbird. Nat Commun 14, 7787. https://doi.org/10.1038/s41467-023-43592-6

Adami, C., 2012. Adaptive walks on the fitness landscape of music. Proceedings of the National Academy of Sciences 109, 11898–11899. https://doi.org/10.1073/pnas.1209301109

Arjmand, H.-A., Hohagen, J., Paton, B., Rickard, N.S., 2017. Emotional Responses to Music: Shifts in Frontal Brain Asymmetry Mark Periods of Musical Change. Front. Psychol. 8. https://doi.org/10.3389/fpsyg.2017.02044

Balaban, H., Ullman, T.D., 2025. The capacity limits of moving objects in the imagination. Nat Commun 16, 5899. https://doi.org/10.1038/s41467-025-61021-8

Baptista, L.F., Keister, R.A., 2005. Why birdsong is sometimes like music. Perspect Biol Med 48, 426–443. https://doi.org/10.1353/pbm.2005.0066

Bedoya, D., Arias, P., Rachman, L., Liuni, M., Canonne, C., Goupil, L., Aucouturier, J.-J., 2021. Even violins can cry: specifically vocal emotional behaviours also drive the perception of emotions in non-vocal music. Philosophical Transactions of the Royal Society B: Biological Sciences 376, 20200396. https://doi.org/10.1098/rstb.2020.0396

Bignardi, G., Wesseldijk, L.W., Mas-Herrero, E., Zatorre, R.J., Ullén, F., Fisher, S.E., Mosing, M.A., 2025. Twin modelling reveals partly distinct genetic pathways to music enjoyment. Nat Commun 16, 2904. https://doi.org/10.1038/s41467-025-58123-8

Brown, S., Savage, P.E., Ko, A.M.-S., Stoneking, M., Ko, Y.-C., Loo, J.-H., Trejaut, J.A., 2014. Correlations in the population structure of music, genes and language. Proceedings

of the Royal Society B: Biological Sciences 281, 20132072. https://doi.org/10.1098/rspb.2013.2072

Ceja, C.R., Franconeri, S.L., 2023. Difficulty limits of visual mental imagery. Cognition 236, 105436. https://doi.org/10.1016/j.cognition.2023.105436

Chernoff, H., 1973. The Use of Faces to Represent Points in k-Dimensional Space Graphically. Journal of the American Statistical Association 68, 361–368. https://doi.org/10.1080/01621459.1973.10482434

Darroch, S.A.F., Smith, E.F., Nelson, L.L., Craffey, M., Schiffbauer, J.D., Laflamme, M., 2023. Causes and consequences of end-Ediacaran extinction: An update. Cambridge Prisms: Extinction 1, e15. https://doi.org/10.1017/ext.2023.12

de Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J Acoust Soc Am 111, 1917–1930. https://doi.org/10.1121/1.1458024

De Gregorio, C., Maiolini, M., Raimondi, T., Carugati, F., Miaretsoa, L., Valente, D., Torti, V., Giacoma, C., Ravignani, A., Gamba, M., 2024. Isochrony as ancestral condition to call and song in a primate. Ann N Y Acad Sci 1537, 41–50. https://doi.org/10.1111/nyas.15151

Ellis, D.P.W., 2007. Beat Tracking by Dynamic Programming. Journal of New Music Research 36, 51–60. https://doi.org/10.1080/09298210701653344

Fitch, W.T., 2015. Four principles of bio-musicology. Philos Trans R Soc Lond B Biol Sci 370, 20140091. https://doi.org/10.1098/rstb.2014.0091

Fitch, W.T., 2006. The biology and evolution of music: a comparative perspective. Cognition 100, 173–215. https://doi.org/10.1016/j.cognition.2005.11.009

Flood, M.W., Grimm, B., 2021. EntropyHub: An open-source toolkit for entropic time series analysis. PLoS One 16, e0259448. https://doi.org/10.1371/journal.pone.0259448

Gardner, A., 2008. The Price equation. Current Biology 18, R198–R202. https://doi.org/10.1016/j.cub.2008.01.005

Gustison, M.L., Bergman, T.J., 2016. Vocal complexity influences female responses to gelada male calls. Sci Rep 6, 19680. https://doi.org/10.1038/srep19680

Gustison, M.L., le Roux, A., Bergman, T.J., 2012. Derived vocalizations of geladas (Theropithecus gelada) and the evolution of vocal complexity in primates. Philos Trans R Soc Lond B Biol Sci 367, 1847–1859. https://doi.org/10.1098/rstb.2011.0218

Harding, E.E., Kim, J.C., Demos, A.P., Roman, I.R., Tichko, P., Palmer, C., Large, E.W., 2025. Musical neurodynamics. Nat. Rev. Neurosci. 26, 293–307. https://doi.org/10.1038/s41583-025-00915-4

Heitmayer, M., 2025. The Second Wave of Attention Economics. Attention as a Universal Symbolic Currency on Social Media and beyond. Interact Comput 37, 18–29. https://doi.org/10.1093/iwc/iwae035

Honing, H., 2018. On the biological basis of musicality. Ann N Y Acad Sci. https://doi.org/10.1111/nyas.13638

Huron, D., 2001. Is Music an Evolutionary Adaptation? Annals of the New York Academy of Sciences 930, 43–61. https://doi.org/10.1111/j.1749-6632.2001.tb05724.x

Izbicki, P., Zaman, A., Stegemöller, E.L., 2020. Music Form but Not Music Experience Modulates Motor Cortical Activity in Response to Novel Music. Front Hum Neurosci 14, 127. https://doi.org/10.3389/fnhum.2020.00127

Jacoby, N., Polak, R., Grahn, J.A., Cameron, D.J., Lee, K.M., Godoy, R., Undurraga, E.A., Huanca, T., Thalwitzer, T., Doumbia, N., Goldberg, D., Margulis, E.H., Wong, P.C.M., Jure, L., Rocamora, M., Fujii, S., Savage, P.E., Ajimi, J., Konno, R., Oishi, S., Jakubowski, K., Holzapfel, A., Mungan, E., Kaya, E., Rao, P., Rohit, M.A., Alladi, S., Tarr, B., Anglada-Tort, M., Harrison, P.M.C., McPherson, M.J., Dolan, S., Durango, A., McDermott, J.H., 2024. Commonality and variation in mental representations of music revealed by a cross-cultural comparison of rhythm priors in 15 countries. Nat Hum Behav 8, 846–877. https://doi.org/10.1038/s41562-023-01800-9

Kepler, J., Aiton, E.J., Duncan, A.M., Field, J.V., 1997. The harmony of the world, Memoirs of the American Philosophical Society. American Philosophical Society, Philadelphia, Pa.

Kim, G., Kim, D.-K., Jeong, H., 2024. Spontaneous emergence of rudimentary music detectors in deep neural networks. Nat Commun 15, 148. https://doi.org/10.1038/s41467-023-44516-0

MacCallum, R.M., Mauch, M., Burt, A., Leroi, A.M., 2012. Evolution of music by public choice. Proceedings of the National Academy of Sciences 109, 12081–12086. https://doi.org/10.1073/pnas.1203182109

Mauch, M., Dixon, S., 2014. PYIN: A fundamental frequency estimator using probabilistic threshold distributions, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 659–663. https://doi.org/10.1109/ICASSP.2014.6853678

Nelson, L.L., Ramezani, J., Almond, J.E., Darroch, S.A.F., Taylor, W.L., Brenner, D.C., Furey, R.P., Turner, M., Smith, E.F., 2022. Pushing the boundary: A calibrated Ediacaran-Cambrian stratigraphic record from the Nama Group in northwestern Republic of South Africa. Earth and Planetary Science Letters 580, 117396. https://doi.org/10.1016/j.epsl.2022.117396

Passmore, S., Wood, A.L.C., Barbieri, C., Shilton, D., Daikoku, H., Atkinson, Q.D., Savage, P.E., 2024. Global musical diversity is largely independent of linguistic and genetic histories. Nat Commun 15, 3964. https://doi.org/10.1038/s41467-024-48113-7

Patel, A.D., 2014. The Evolutionary Biology of Musical Rhythm: Was Darwin Wrong? PLoS Biol 12, e1001821. https://doi.org/10.1371/journal.pbio.1001821

Peretz, I., Vuvan, D., Lagrois, M.-É., Armony, J.L., 2015. Neural overlap in processing music and speech. Philosophical Transactions of the Royal Society B: Biological Sciences 370, 20140090. https://doi.org/10.1098/rstb.2014.0090

Pfattheicher, S., Nielsen, Y.A., Thielmann, I., 2022. Prosocial behavior and altruism: A review of concepts and definitions. Current Opinion in Psychology 44, 124–129. https://doi.org/10.1016/j.copsyc.2021.08.021

Rafii, Z., 2022. The Constant-Q Harmonic Coefficients: A timbre feature designed for music signals [Lecture Notes]. IEEE Signal Processing Magazine 39, 90–96. https://doi.org/10.1109/MSP.2021.3138870

Raimondi, T., Di Panfilo, G., Pasquali, M., Zarantonello, M., Favaro, L., Savini, T., Gamba, M., Ravignani, A., 2023. Isochrony and rhythmic interaction in ape duetting. Proc Biol Sci 290, 20222244. https://doi.org/10.1098/rspb.2022.2244

Ravignani, A., Bowling, D.L., Fitch, W.T., 2014. Chorusing, synchrony, and the evolutionary functions of rhythm. Front Psychol 5, 1118. https://doi.org/10.3389/fpsyg.2014.01118

Richman, B., 1987. Rhythm and melody in gelada vocal exchanges. Primates 28, 199–223. https://doi.org/10.1007/BF02382570

Rothenberg, D., Roeske, T.C., Voss, H.U., Naguib, M., Tchernichovski, O., 2014. Investigation of musicality in birdsong. Hear Res 308, 71–83. https://doi.org/10.1016/j.heares.2013.08.016

Sacks, O., 2008. Musicophilia: Tales of Music and the Brain, Revised and Expanded Edition, Revised&enlarged edition. ed. Vintage, New York.

Savage, P.E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., Fitch, W.T., 2021. Music as a coevolved system for social bonding. Behavioral and Brain Sciences 44, e59. https://doi.org/10.1017/S0140525X20000333

Sawant, S., Arvind, C., Joshi, V., Robin, V.V., 2022. Spectrogram cross-correlation can be used to measure the complexity of bird vocalizations. Methods in Ecology and Evolution 13, 459–472. https://doi.org/10.1111/2041-210X.13765

Snowdon, C.T., Teie, D., 2009. Affective responses in tamarins elicited by species-specific music. Biology Letters 6, 30–32. https://doi.org/10.1098/rsbl.2009.0593

Spectral Processing - DAFX - Wiley Online Library [WWW Document], n.d. URL https://onlinelibrary.wiley.com/doi/10.1002/047085863X.ch10 (accessed 9.7.25).

Spitzer, M., 2021. The Musical Human: A History of Life on Earth. Bloomsbury Publishing USA, New York.

Trimble, M., Hesdorffer, D., 2017. Music and the brain: the neuroscience of music and musical appreciation. BJPsych Int 14, 28–31.

Trost, W., Trevor, C., Fernandez, N., Steiner, F., Frühholz, S., 2024. Live music stimulates the affective brain and emotionally entrains listeners in real time. Proc Natl Acad Sci U S A 121, e2316306121. https://doi.org/10.1073/pnas.2316306121

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Vettehen, P.H., Schaap, G., 2023. An attention economic perspective on the future of the information age. Futures 153, 103243. https://doi.org/10.1016/j.futures.2023.103243

Vuust, P., Heggli, O.A., Friston, K.J., Kringelbach, M.L., 2022. Music in the brain. Nat Rev Neurosci 23, 287–305. https://doi.org/10.1038/s41583-022-00578-5

Watkins, H., 2018. Musical Vitalities: Ventures in a Biotic Aesthetics of Music, New Material Histories of Music. University of Chicago Press, Chicago, IL.

Wigner, E.P., 1960. The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959. Communications on Pure and Applied Mathematics 13, 1–14. https://doi.org/10.1002/cpa.3160130102

Wolfe, J., 2002. Speech and Music, Acoustics and Coding, and What Music Might Be “For.”

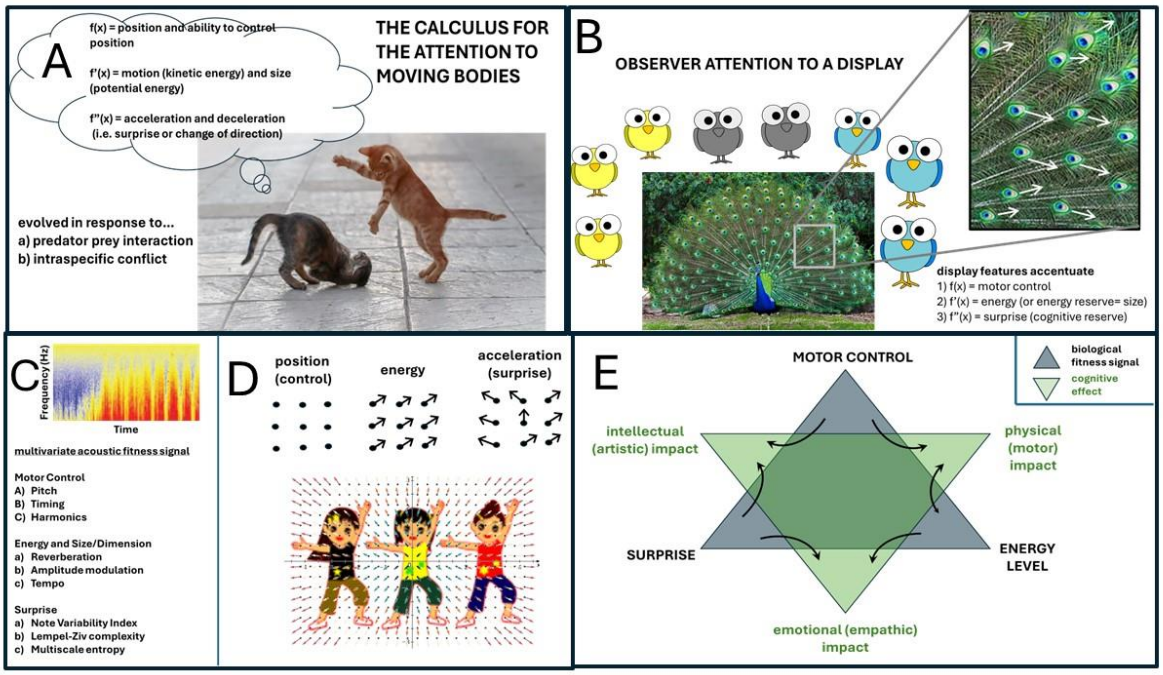

Figure 1. The theory of the calculus of attention examined through musical stimulation of human minds. *(A) The calculus of attention focuses upon three components of biological fitness; motor control, energy level, and cognitive capacity demonstrated via surprise or novelty. (B) It is captured by many proximate stimuli present in animal displays. For example, a female bird might choose among prospective males based upon how loud or fast they sing (indicator of energy), how well they conform to a species-specific call (indicator of motor control), and in some species such as mockingbirds, nightingales, and lyrebirds, how well they can learn to improvise (indicator of cognitive reserve). (C) Human music and dance also capture attention using these same three elements of display. (D) These outward displays of biological fitness signaling by an individual musician or group of musicians (upright gray triangle) combine to affect the minds of performers and listeners by stimulating physical, emotional, and intellectual sensations (inverted green triangle).*

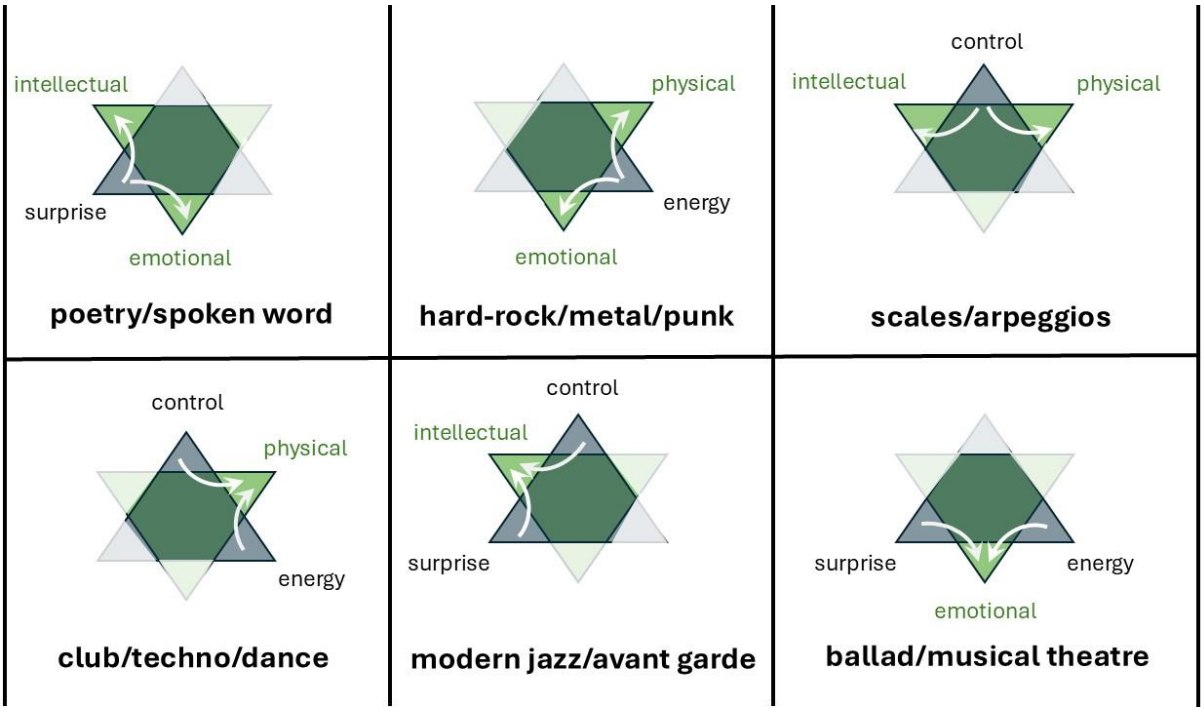

Figure 2. The effects of increased emphasis on one or two proximate stimuli in the calculus of attention (gray) can explain the cognitive reactions (green) to various artistic genres through their adjacency in the geometric pattern*. For example, emphasis on just energy level in music with de-emphasis on novelty/surprise and fine motor control can result in the strong feelings of physical plus emotional impact that we often tend to associate with heavy metal/punk rock. Similarly, a singular emphasis on motor control produces a consequent lack of emotion when we listen to scales or arpeggios, and a singular focus on lyrical novelty, without emphasis energy or motor control explains the cognitive differences we perceive when comparing poetry/spoken word versus music itself. When adjacent pairs of fitness signals are emphasized in combination, it has the effect of focusing on the cognitive impact in a single area. For example, the primarily physical impact we hear in dance music is created by a combination of balanced emphasis on energy level and motor control, with little novelty or surprise. Similarly, the intellectual stimulation of modern jazz is created by the balanced combination of improvisation (surprise) with motor control. And the enhanced emotional impact of song ballads and musical theatre are caused by the balance of energy level and unexpected features in the lyrical or musical elements. Note: not pictured here, popular music would seem to require musicians to maintain balance between all three features of the model (energy, control and surprise).*

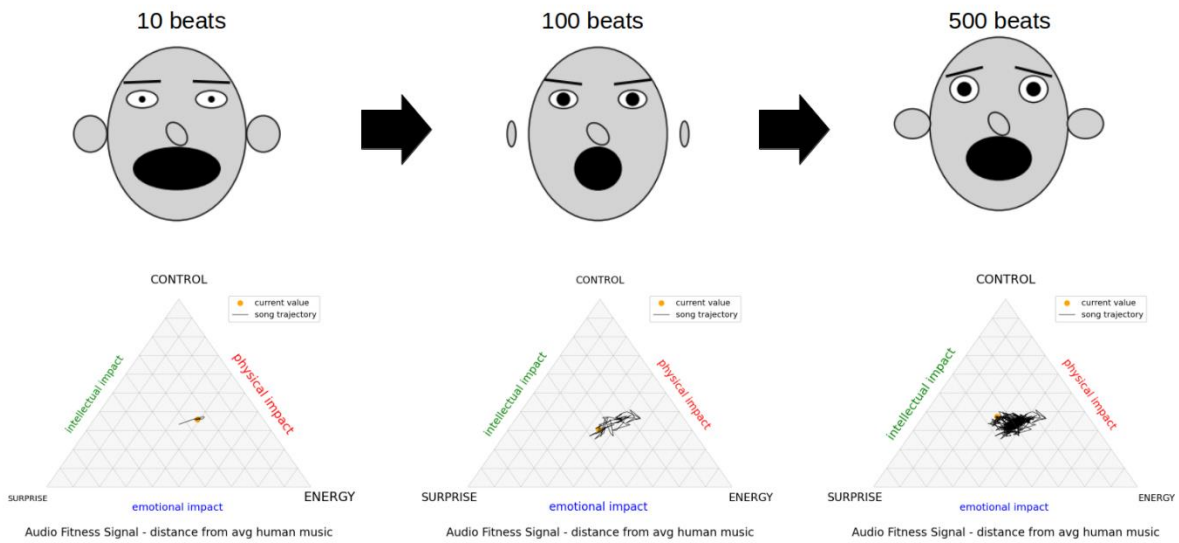

Figure 3. The POPSTAR software *measures three indicators each for sound control (pitch control, timing evenness, and harmonic levels), sound energy (tempo, amplitude modulation, and reverberation), and surprise or sound complexity (note variability index, Lempel-Ziv complexity, multi-scale entropy). It outputs a movie (.mp4) file for each song track as a trajectory within a dynamic ternary or triangle plot in the combined space of control, energy, and surprise (CES) as well as each feature represented by the components of a dynamic Chernoff face, where ear/nose dimensions represent control, mouth/eyebrow dimensions represent energy and eye dimensions represent surprise. The movie is updated according to the user, either on every time beat in the music (determined by beat tracking algorithm) or every ¼ of a second. The fitness signaling in musical sounds are determined through further analysis of the time step distribution of the song trajectory in the ternary plot.*

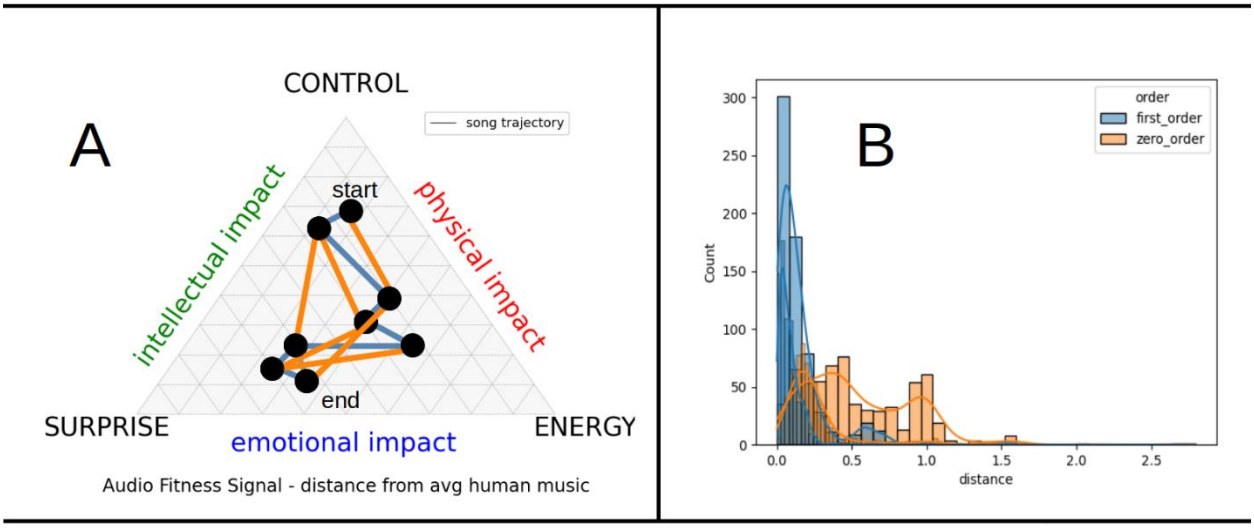

Figure 4. The behavioral fitness signal in the space of sound control, energy, and surprise (CES) *is determined by comparing the step distribution of the song trajectory in the space of CES (blue) to random walks of the song trajectory created by the shuffling of the order of steps (yellow). An expertly created musical sound typically shows a lognormal, gamma or power lognormal distribution of steps where step distances are usually very short and controlled until the musician intends to move somewhere else in the space of CES. When the time dependence of the signal is destroyed by shuffling, the distribution of steps becomes more uniform with a larger mean step distance. In the case of brown noise, no difference in observed and shuffled step distributions is evident. The percentage of the stability in CES can be determined through repeated permutations of random draws from the real- and shuffled-time step distributions and tallying the number of times the time step in the observed trajectory is less than the that of the shuffled trajectory. The significance of this difference can be determined via the Kruskal-Wallis test.*

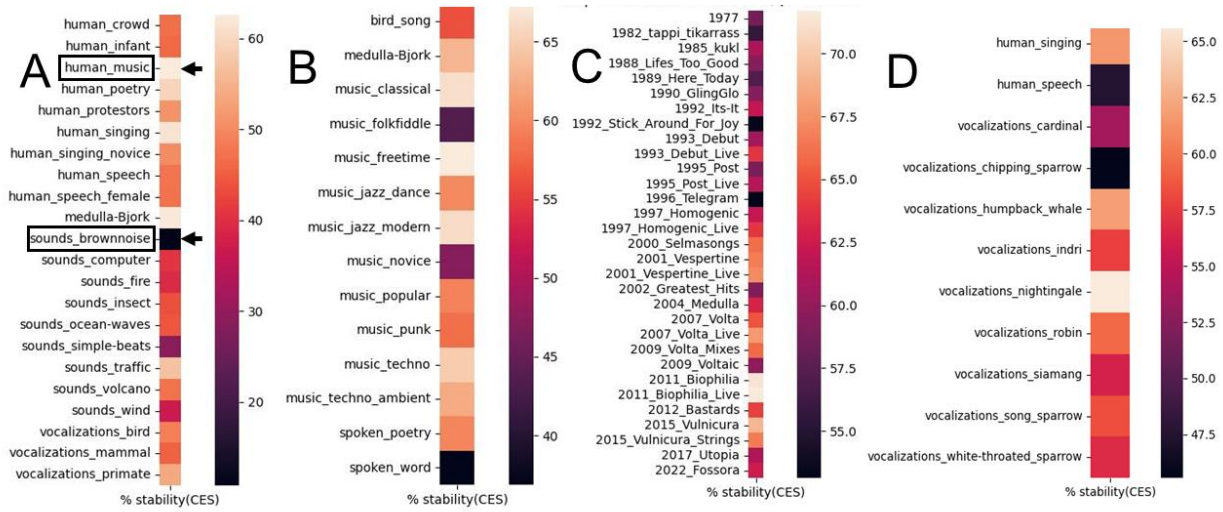

Figure 5. Relative stability of sound trajectories in the space of control, energy, and surprise (CES) *for (A) selected natural sounds, (B) musical genres, (C) a pop star's career discography and (D) selected bird song and primate vocalizations. Note that (A) noise has 0% stability in the space of CES while music has the most (typically > 50%) which is (B) roughly 10% higher than spoken words. Also note (D) that in this measure, birds with complex and melodic songs have more CES stability than animals with less musical sounding vocalization.*

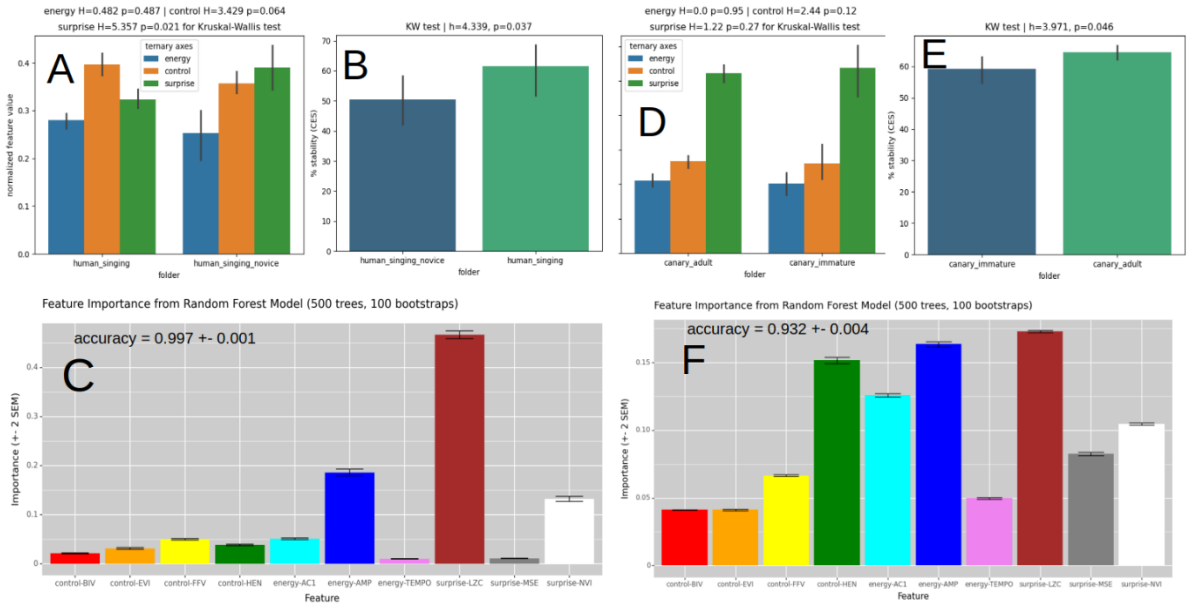

Figure 6. Comparison of stability in musical control, energy and surprise (CES) in expert vs. novice singers (A-C) and adult vs. immature birds (D-F). *(A) Expert singers exhibit significantly more control and less surprise or complexity than novice singers. (B) Expert singers also have significantly more stability in song trajectory in the space of CES than novice singers. (C) Random forest classification indicates with high accuracy that differences between experts and novices are largely driven by Lempel-Ziv (substring) complexity, amplitude modulation, and note variability in this study.*

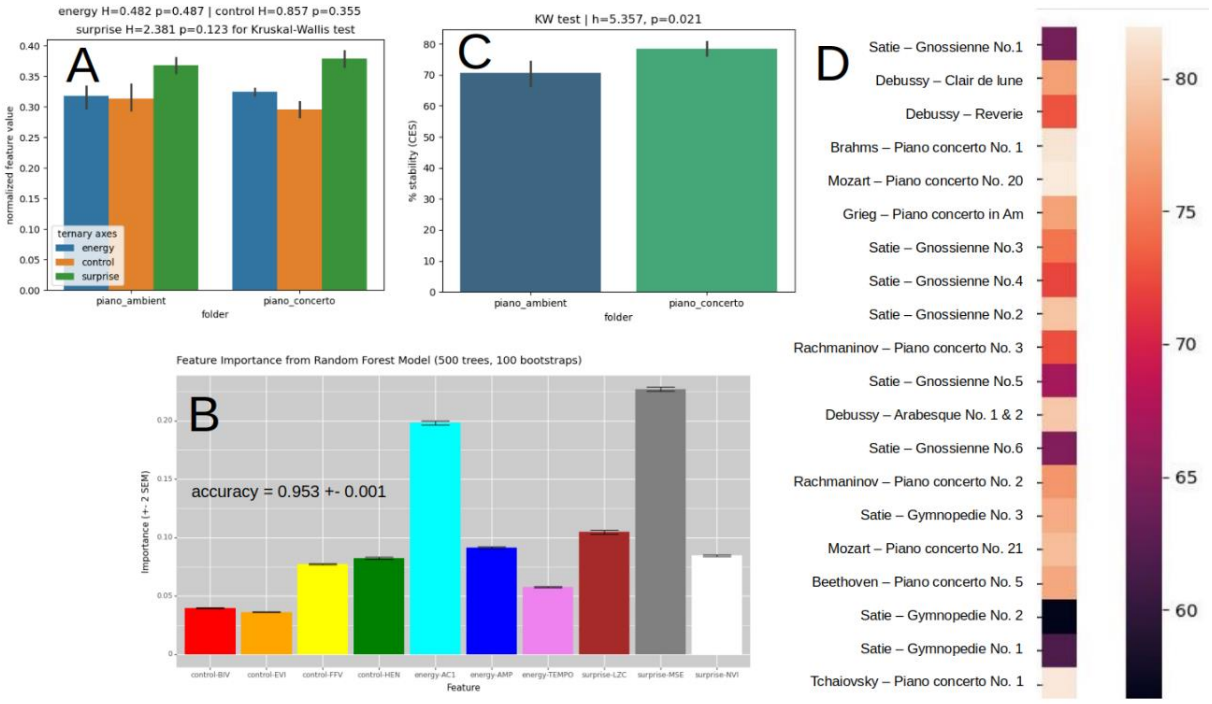

Figure 7. Comparison of stability in musical control, energy and surprise (CES) in piano concertos vs. ambient piano pieces (i.e. furniture music of Erik Satie and Claude Debussy). *(A) While there are no significant differences in the levels of CES between the two genres, (B) the piano concertos have significantly more stability in song trajectory in the space of CES than ambient music. (C) Random forest classification indicates with high accuracy that differences between the two genres are largely driven by multiscale entropy and reverberation.*

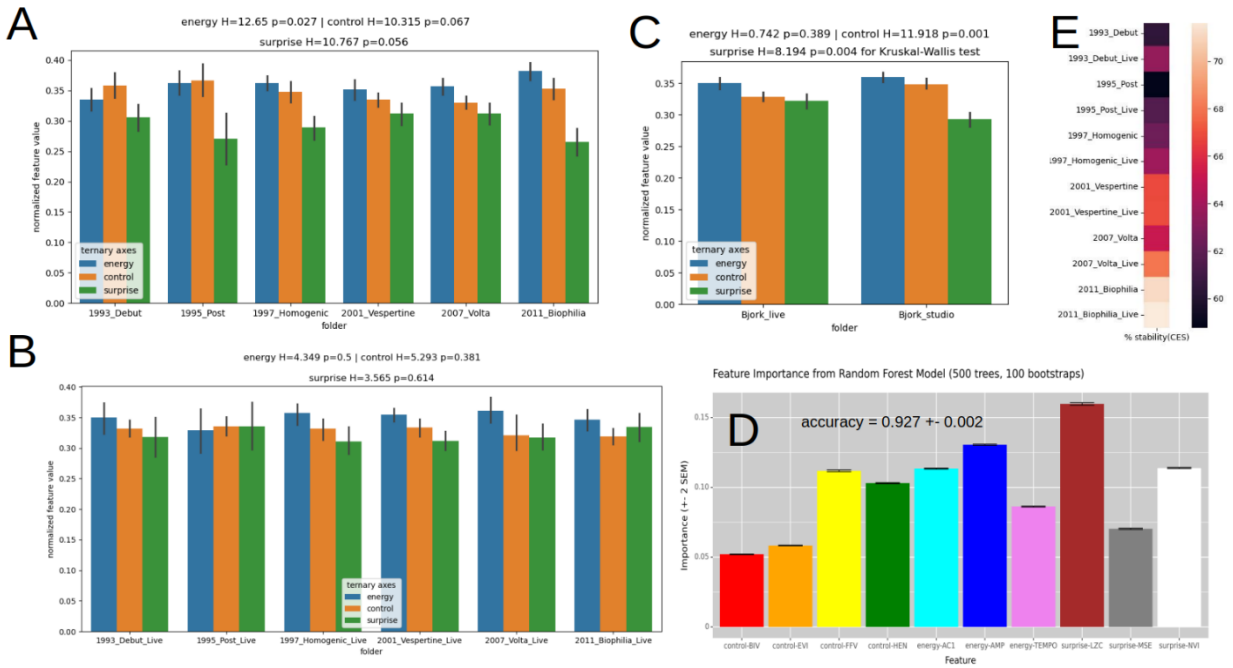

Figure 8. Comparison of stability in musical control, energy and surprise (CES) in live vs. studio performances while controlling for individual song track. Here we examine the Icelandic pop star Bjork's six studio LPs compared to their corresponding live tour LPs. We find that (A) the studio LP's exhibit significant or borderline significant differences in CES while (B) the corresponding live LP's do not. This suggests that live performances feedback may constrain Bjork's performance to a more consistent level of relative CES. (C) Overall, the live performances show significantly less control and more surprise than the corresponding studio song tracks, showing a trend similar to that distinguishing novice from expert performance. This suggests that fitness signaling via CES is less manipulatable on stage than in studio. (D) *Random forest classification indicates with moderate accuracy that differences between Bjork's live and studio performance involves all nine features of CES to large extent. (E) While each studio project and its accompanying tour vary significantly in CES stability, the live performances seem to be slightly more stable in CES, again suggesting some creative constraints or added fitness requirements when performing live.*

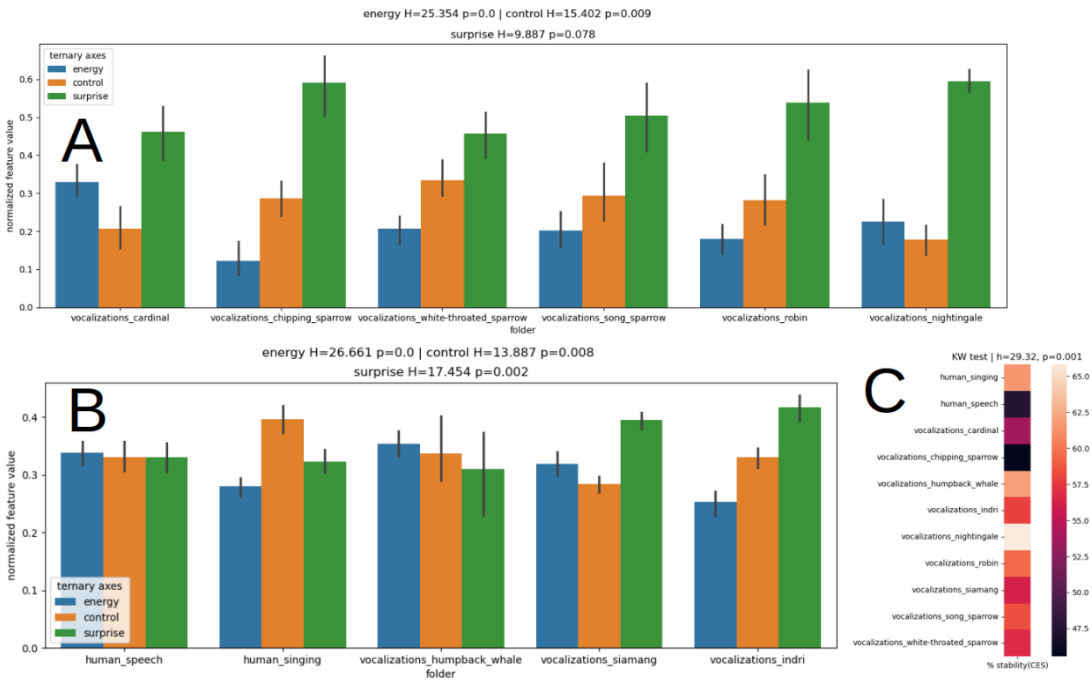

Figure S1. Species comparison in vocal control, energy and surprise (CES) across (A) birds and (B) primates and whales. (C) Comparative stability in CES is also shown.

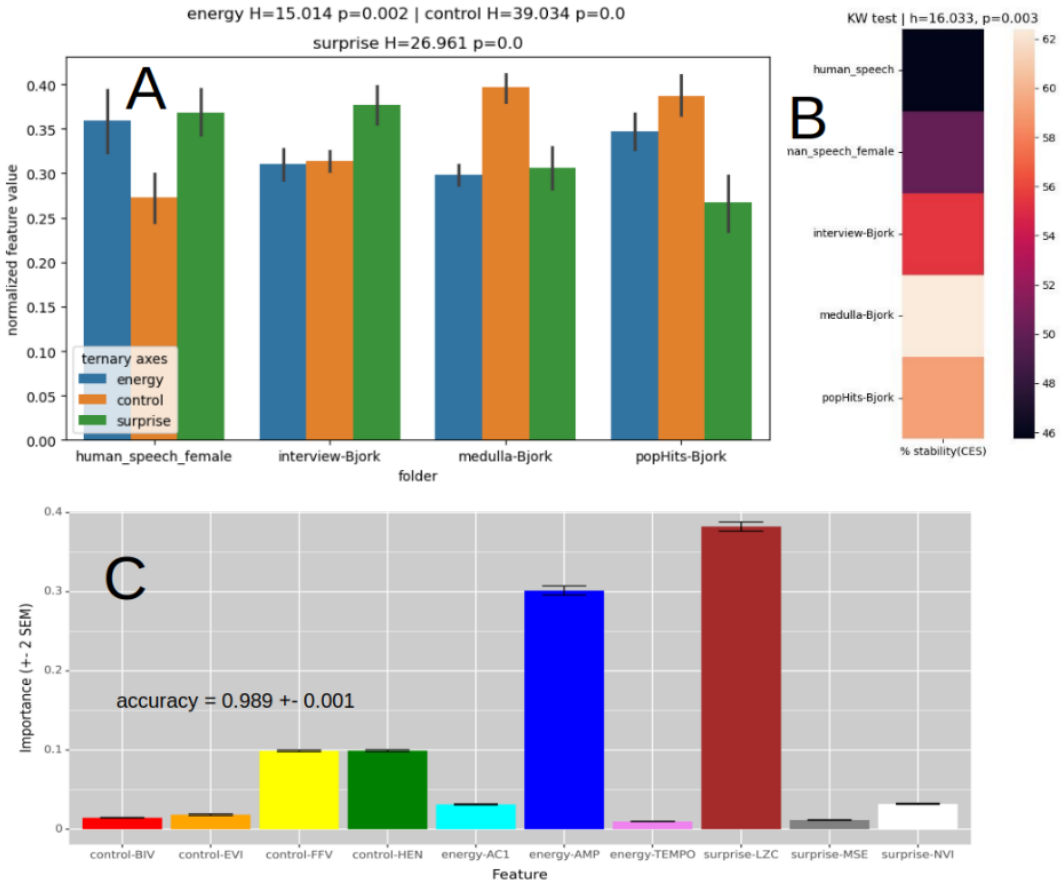

Figure S2. Comparison of stability in musical control, energy and surprise (CES) in natural speech, music with vocals and instruments, and music created solely with human vocals controlled for a single performer's voice. *Here, the Icelandic pop star Bjork is analyzed during her interviews, her greatest hits LP, and her 5th studio album Medulla, which is derived almost entirely with human vocals emulating modern instruments. We see that music has (A) more control and less surprise than speech and (B) more stability in CES than speech, suggesting that the results in Figure 6 could be due to novice singers over-emphasizing patterns normal to speech in their attempts to sing. (C) Random forest classification indicates with high accuracy that differences between Bjork's speech and music are largely driven by Lempel-Ziv (substring) complexity, amplitude modulation, and to a lesser extent, control over pitch and timing.*