In-depth analysis of music structure as a text network

Ping-Rui Tsai,[1] Yen-Ting Chou,[1] Nathan-Christopher Wang,[2] Hui-Ling Chen,[3] Hong-Yue Huang,[1] Zih-Jia Luo,[4] and Tzay-Ming Hong[1,*]

[1]*Department of Physics, National Tsing Hua University, Hsinchu 30013, Taiwan, Republic of China*

[2]*College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109, USA*

[3]*Department of Chinese Literature, National Tsing Hua University, Hsinchu 30013, Taiwan, Republic of China*

[4]*Advanced Semiconductor Engineering, Inc., Kaohsiung 76027628, Taiwan, Republic of China*

(Received 10 March 2023; accepted 8 August 2024; published 11 September 2024)

Music, enchanting and poetic, permeates every corner of human civilization. Although music is not unfamiliar to people, our understanding of its essence remains limited, and there is still no universally accepted scientific description. This is primarily due to music being regarded as a product of reason and emotion, making it difficult to define. This article treats musical texts as a complex system. This view echoes linguist John Rupert Firth's insight that understanding a word involves defining it through its surrounding relationships. To construct the network we first build a linear regression model with threshold values to assign conditions to the links among note, time, and volume. Then a clustering coefficient representing regional characteristics is utilized to define the word. Finally, the statistical distribution of the text is strictly required to adhere to the grammatical properties of statistical linguistics, such as Zipf's law, to adjust the weights of the linear regression model and achieve optimal results. These processes enable us to comprehend the structural differences in music across different periods with scientific rigor. Relying on the advantages of structuralism, we concentrate on the relationships and order between the physical elements of music, rather than getting entangled in the blurred boundaries of science and philosophy. Aside from serving as a bridge connecting music to natural language processing and knowledge graphs, the technical methods developed in this work offer a more intuitive approach to elucidate the relationships among elements of a complex network.

DOI: 10.1103/PhysRevResearch.6.033279

I. INTRODUCTION

Music is often considered a form of natural language [1,2] because it exhibits the ability to adapt [3] and coevolve with human civilization. According to Soviet musicologist Genrikh "Henry" Orlov, music can participate in communication and unite people through a single emotion. The characteristics that make each language unique may be adaptations to the acoustics of different environments [4]. In addition, the development of languages and music can be categorized into distinct stages that reflect the historical events of each period, and both originate from the imitation of sounds from the environment; the influence of the environment on the variations in vocalization among organisms has also been confirmed [5].

Music, like language, has developed its own notation and system of reading and writing. The meaning conveyed by music is less precise than spoken language due to its lack of tenor and vehicle [6,7]. Despite this weakness, music is no less capable of evoking our memories of specific experiences [8]. By using magnetoencephalography (MEG), it has been confirmed that music and language are governed by the same mechanism in the cerebral cortices. This implies the similarity in their process of data that are transmitted through the sounds associated with spoken language and music harmony [9]. The research on the joint precursor of language and music, concerning how they are constructed by sound, is related to their evolution from a common origin and later divergence [10].

Let us proceed to compare their structure. Both involve minimal units, such as words and chunking, that serve as the building blocks to construct the corpus and scores through ever larger hierarchical units, i.e., phrases/sentences and the deep structure [11,12]. However, exploring music subsystems such as melody, harmony, and counterpoint may pose challenges due to their complex interrelations in the music syntax. Similarly, the varying ratios of morphemes to each word in morphological typology result in distinctions between isolating and polysynthetic languages. How this difference impacts the definition and boundaries of words is an ongoing issue in linguistics. For example, fusional languages [13] employ irregular methods of word formation or combine morphemes from multiple concepts, thus making it challenging at times to discern the original morphological relationships, while agglutinative languages like Turkish construct words by cohesively affixing multiple morphemes [14].

With the development of statistical linguistics, Zipf [15,16] empirically found that the frequency-rank distribution of

*Contact author: ming@phys.nthu.edu.tw

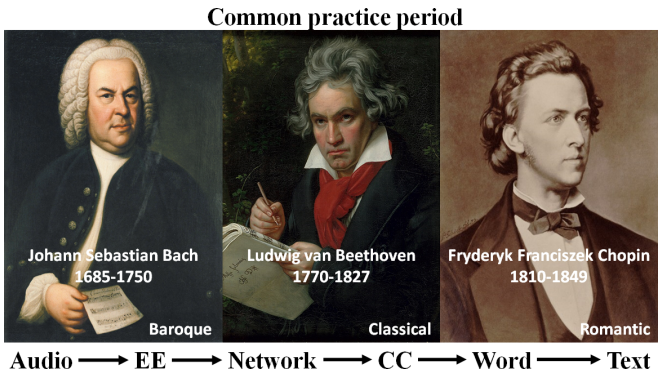

FIG. 1. This illustration depicts the three main periods we primarily discuss in the article: the Baroque, Classical, and Romantic periods, each represented by a representative composer. Additionally, we visually present the overall processes of the textualization of music in essential elements (EEs).

corpus and natural language utterances follows the power law $y = a/x^b$ where a, b are constants. This distribution was later established to be prevalent in the ranking of many natural and manmade systems, such as web links [17] and brain functional networks [18,19]. Although Zipf's law has been confirmed to exist in musical composition, there is no consensus on what unit plays the role of a word. For example, it has been proposed via segmentation aided by the application of equal temperament to pitch, timbre, and loudness or through binary coding on the power spectrum [20,21]. The numerous attempts above, regardless of the way word definitions are approached, all revolve around definitions based on physical categories.

In this article, we deconstruct fundamental musical elements like rhythm, timbre, pitch, melody, articulation, meter, and tempo, and then transform them into concepts derived from physics—specifically, space, time, and volume. These three elements are herein referred to as essential elements (EEs). Notes ($C^1, D^1, E^{\flat 1}, E^1, F^{2,}, G^2$, etc.) and time, along with its corresponding beat, are important representations of the universal features of music [22]. Employing a reference coordinate system based on 0.1-sec intervals and notes on a piano, we establish linking conditions between each pixel (node) to form an evolutionary network. Additionally, we define words by clustering coefficients (CC)

$$CC(i) \equiv \frac{2 \times E(i)}{D(i) \times (D(i) - 1)}, \tag{1}$$

where $E(i)$ denotes the number of actual edges between the neighbors of node i, and $D(i)$ is the degree of node i. The processes involved in generating the word out of a musical audio file are described schematically in Fig. 1. The main reason for choosing CC instead of the degree is the former's ability to reflect the relationships among the nodes within each region, in addition to describing those between each node and the neighboring nodes connected to it.

Similar to contemporary large language models (LLMs), our text generation process relies on hierarchical relationships. LLMs utilize an extensive vocabulary table encompassing various lexical items and tags, including verbs, nouns, adjectives, and others—all seamlessly integrated into the model. Each word is associated with an embedding vector learned during the training. Contextual clues interact with this table, determining the similarity between the contextual encoding vector and each embedding vector. This similarity also reflects the likelihood of each word within the given context [23].

Here we will answer five questions: (1) How does one generate text from music using all of the essential elements in music? (2) What are the linking conditions of music across different periods in an evolutionary network and the variations under Zipf's law distribution? (3) How does the diversity in the choice and frequency of words within an evolutionary network across different periods reflect the versatility in song structures? (4) How robust and how free is the network that reflects the characteristic features of each music periods against random removal of words? (5) How do we use an audio structure to discern music from nonmusic? Finally, how can we observe the evolution and extinction of musical words akin to the evolution of natural language by tracking the magnitude of CC?

II. NETWORK MODELING

To answer the first question, which is inspired by previous methodologies in network science for generating specific network structures [24], we establish linking conditions based on interactions among EEs. Subsequently, these conditions can be used to determine the network structure. We quantify the variations between these EEs and establish a threshold for the validation of links to form a structure. The music signals can then be analyzed in the frequency and time-frequency domains, with a frequency range of 1–8192 Hz and a time increment of 0.1 sec. Additionally, we will transform these domains into the note-time domain. The note domain consists of 84 tones or pitches, which correspond to those of a piano in equal temperament and cover a frequency range of 1–8192 Hz. Expressing the volume in normalized decibels from 0 to 10, we eliminate the volume that is less than 0.1, based on the power-time spectrum in order to reduce the difference in volume that was derived from the recording process.

We then define the amount of information (I) via comparing the change of note position (N), time position (T), and volume (V) of pixels 1 and 2 in note vs time coordinate:

$$I \equiv w_1|\mathrm{N}(1) - \mathrm{N}(2)| + w_2|\mathrm{T}(1) - \mathrm{T}(2)| + w_3|\mathrm{V}(1) - \mathrm{V}(2)| + w_4|\mathrm{V}(1) + \mathrm{V}(2)|. \tag{2}$$

How these weights $w_{1,2,3,4}$ are chosen will be explained shortly. Via the simple linear regression in Eq. (2), the weights can be determined and shown to offer a high level of interpretability of the musical form across these four aspects when composing. The first three terms in Eq. (2) reflect the musical form of the composition, while the fourth is an additional term that represents the energy carried by the intentions and emotions that the composer tried to convey [25]. The amount of information carried by any music is limited and determined by both the composer's style and intentions [26]. We rarely see dramatic changes in all elements in musical pieces, and, therefore, it is expected that there exists an upper threshold I_{m} for the amount of information that can be

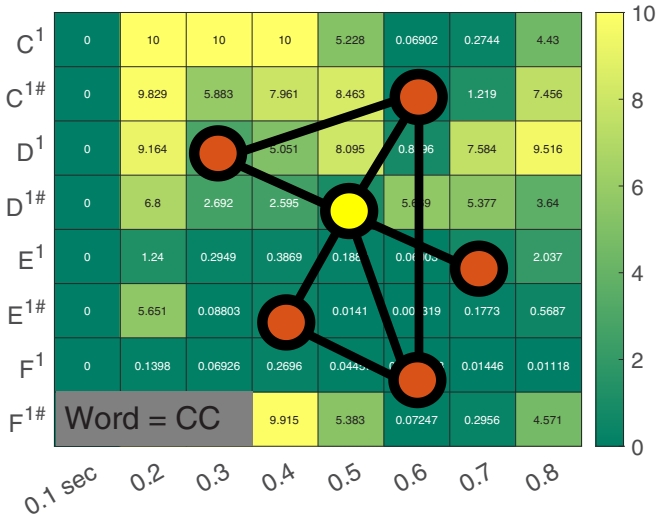

FIG. 2. Volumes are stated in the note vs time plot for the music score by Ryuichi Sakamoto. The exchange of information between two pixels is determined by their elements and weights. Using Eq. (2), we can calculate the linking condition for the text network of the music and define the link between dots. The *CC*s in the network are then treated as words. Take the exemplary network in this figure, for instance. The $D = 5$ and $E = 3$ so that $CC = (2\times3)/(5\times4) = 0.3$, according to Eq. (1).

conveyed between pixels in the local interaction:

$$\text{Link} = \begin{cases} 1 & \text{if } I < I_{\text{m}}, \\ 0 & \text{if } I > I_{\text{m}}. \end{cases}$$

This also implies that, within limited variations, the priorities considered by the composer determine the style of composition.

Using the total number of links, we can calculate $\text{CC} = 2N/[K(K-1)]$ of each pixel, where the degree N counts the number of neighbors and K the number of links among them. In order for CC to be interpreted as a word, it has to be significant enough at supplying the semantic meaning of the cluster based on the quantities N and K. A sample result of the above procedures can be found in Fig. 2. This definition of word not only is consistent with the structuralism of musical material [27], but also follows the spirit of natural language modeling [28,29], such as linguist John Rupert's interpretation of "You shall know a word by the company it keeps" [30]. This spirit expresses the role of endowing meaning to a word via its interaction with neighbors.

Having defined the equivalent of words in music is not enough. We still need to make sure that the structure derived from them obeys the statistical properties in linguistics, notably the empirical Zipf's law. This imposes a constraint that can be utilized to select the most suitable weights w and I_{m} and threshold that give the best fitting by a power-law distribution for word frequency vs ranking. Finally, the w and I_{m} function like a fingerprint that is unique to the structure and helpful for us to distinguish the music from different periods. We can also use them to compare the properties of musical and nonmusical structures; for example, as the composer Edgar Varése said, "Music is organized sound" [31].

We need to select the optimal weight combination from 4032 configurations to accurately represent the text representation of the music. Weights set the range for note and time weights to be 0.05–0.3 with an interval of 0.05 and the range for volume to be 0.1–0.4 with an interval of 0.1. We also normalized the decibel values between 0 and 10 for all data points. The threshold range was set to be 0.5–1.8 with an interval of 0.2. In order to reduce the computing time, we will focus only on the neighboring nodes that are separated within seven notes and 0.7 sec.

The optimal values for the weights in Eq. (2) were determined by two criteria: first, the distribution of Zipf's law for CC in the music score must exhibit an R-square value exceeding 0.8 after deleting the first rank and plotting the frequency vs rank in full logarithm. Second, the largest type of CC should be selected as the optimization condition in order to extract the maximum number of word types to maintain diversity.

From now on, we shall name our evolutionary network the Essential Element Network (EEN) to emphasize its inclusion of essential EEs in music. Our analyses focus on piano pieces from the Common Practice Period (CPP), which spans from 1650 to 1900 [32]. The CPP marks the establishment of the Western musical system and the definition of the harmonic system, which is essential for the interaction between musical elements [33,34] in that period. Within the CPP, the Baroque period (B) is the earliest, followed by the period of Classical music as exemplified by Beethoven (BN), and the Romantic period (R) is the most recent in this time frame.

III. DISTINGUISHING DIFFERENT MUSICAL PERIODS

A. Criterion: Weights

To answer question 2, we can examine the weights of music from different periods to understand the variations in network structure. We used the t-distributed stochastic neighbor embedding (t-SNE) as a dimension reduction method to visualize the distribution of weights in the three different musical periods [35]. The clustering of data points for music from the Classical and Baroque periods in Fig. 3(a) implies their weights are more similar to each other than to the Romantic. In Fig. 3(b) it can be seen that during the Baroque period, w_3 is the most prominent weight, which is related to the performance of clear gaps between notes, with the aim of maintaining clear voices.

The w_1 and w_2 are chosen with similar weights, representing the characteristics of Baroque polyphonic music, emphasizing rigorous counterpoint [36]. However, the priority of weights shifts in the Romantic period when people tried to break free from the constraints of composition, leading to a greater freedom and evolution of more diverse composition styles [37,38]. A t-test analysis of the weights shows that there exists a significant distinction between w_1 and w_3 if the P value is less than 0.001. The combination of weights behind the EEN reflects the characteristic style of composition in each period by use of Zipf's law, as shown in Fig. 3(c). In Fig. 3(d) we also compare different types of audio to test the applicability in 4032 combinations of weights. It turns out that Morse code did not follow the expected power-law

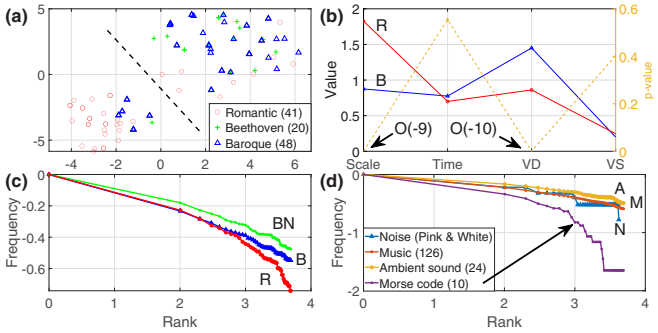

FIG. 3. (a) A t-SNE mapping of the four weights and threshold value onto the eigenspace. The dashed line is to highlight the existence of two clusters, as indicated by the statistical population in parentheses. (b) A t test is conducted to assess the statistical significance of the weight selection where the orange dotted line is for the P value on the right y axis. (c) This full logarithmic plot for the Zipf distribution in different periods. (d) The Zipf distribution of different types of sounds under the range of weight selection, We use the initial letter to represent a line, in which ambient sound includes bird, river, and city traffic.

distribution, which suggests that the chosen weights are not suitable.

B. Criterion: Trend and histogram of words

Here we will answer question 3. After CC is calculated by scanning from the first to last notes, the resulting sequence is found to be periodic in Figs. 4(a) and 4(b). Based on the sequence, we can analyze the representation of words. Figure 4(c) shows that the Baroque period, which emphasized musical formalism, has a uniform development in performance. In contrast, the Romantic period adopted a strategy of destroying or escaping the previous musical forms. Ambient sound is shown to share a similar distribution to the Baroque period, but with more fluctuations and different CC

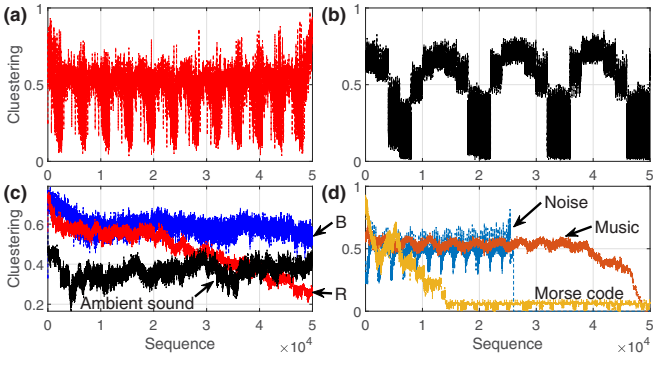

FIG. 4. The CC is plotted against sequence. Two samples of EEN distributions in one dimension are shown in (a, b), both of which turn out to be periodic. (c) Compared to the Baroque period and ambient sound, the variation of CC in the Romantic period is more pronounced, which is in line with the conventional view in musical history that its musical form is more diverse. (d) The distribution of different types of audios where the music samples include 38 piano pieces with a length of less than 2 min.

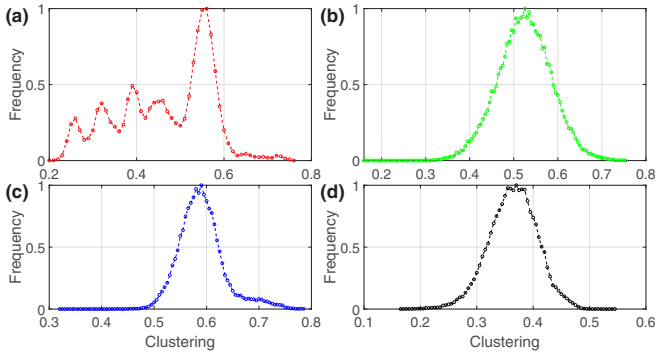

FIG. 5. Histogram of the Romantic, Classical, and Baroque periods and the ambient sounds is shown respectively in (a), (b), (c), and (d). The Romantic period is the only one that defies the normal or Gaussian distribution at exhibiting multiple peaks.

intervals. In Fig. 4(d) we see that the distribution of CC for music resembles that of white and pink noises, in contrast to that of Morse code. This is reasonable because music often contains elements that are abstract and hard to assigned any meaning, whereas the tenor and vehicle of Morse code are always precise.

Having shown the trend of CC, let us now analyze their histogram in Fig. 5. The normal distribution is obtained for the ambient sound as well as all musical periods, except the Romantic one, which exhibits multiple peaks which reflect more diversity in its selection of words.

IV. DEEP LEARNING ANALYSES

A. Training information

To identify the unique characteristics of CPP, we use word mapping to represent CPP by a 2D EEN. This is done by filling the CCs to their corresponding pixel in the note vs time plot of Fig. 2, and applying a convolutional neural network (CNN) [39] to classify musical periods. CNN is a powerful image classification model that can predict the label of an image, relying on its ability to extract local spatial features and keep translation invariant.

To test the robustness and minimum discernible information carried by a text, we design the following two tasks: (1) determine the minimum size of text to represent each of the three CPP periods and (2) test the rigor of conventional definition for these periods. Treating the text as a network, we follow the conventional approach of random node removal to explore the robustness of network features. Instead of randomly deleting network nodes [40], we remove words from 2D EEN to understand the maximum amount of disruption each CPP can withstand and be effectively recognized by a CNN. We expect this knowledge can be used to infer the relative amount of rules in musical form. Strictness implies more rules that will likely render the network more vulnerable to disruptions. More details of the procedure and results of training shown, respectively, in Figs. 5 and 6 will be discussed in Secs. IV B and IV C.

Due to the numerous CNNs used in Fig. 7(a), we provide only the minimum and maximum amount of 2D EEN for the learning curves in Figs. 6(a) and 6(b) for task 1 with

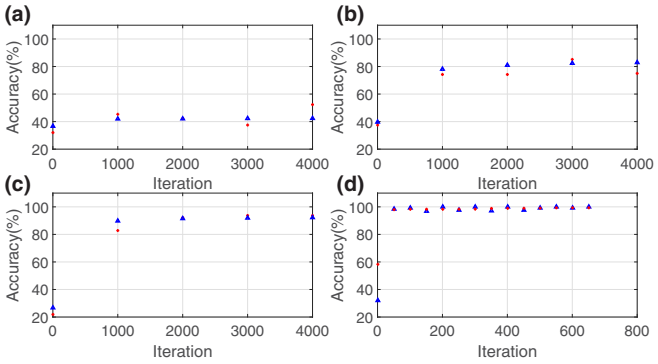

FIG. 6. Panels (a), (b), and (c) represent the learning curves for minimum text size to classify three CPP periods for texts with 2 notes, 0.3 sec, 72 notes, 3.5 sec, and 84 notes, 10.1 sec where the blue/red line denotes the training/validation curve. Panel (d) shows similar curves for classifying music or nonmusic with 84 notes, 10.1 sec.

2 notes/0.3 sec and 72 notes/3.5 sec from 34 000 and 32 456 samples. Similarly, we need to train the CNN to obtain a test accuracy exceeding 93% for the prediction of the 2D EEN's period in preparation for task 2. The result is shown in Fig. 5(c). Figure 6(d) is dedicated to distinguishing music and nonmusic with a test accuracy equaling 100%. Both Figs. 6(c) and 6(d) adopt 84 notes/10.1 sec and are trained from 12 000 samples. So far, the CNN adopts four convolutional layers of 2×2 kernel size and three fully connected layers with 128, 64, and 3 nodes. The hyperparameters are a training-validation-test ratio of 6:2:2, batch size of 64, epoch of 6, and a learning rate of 10^{-4}.

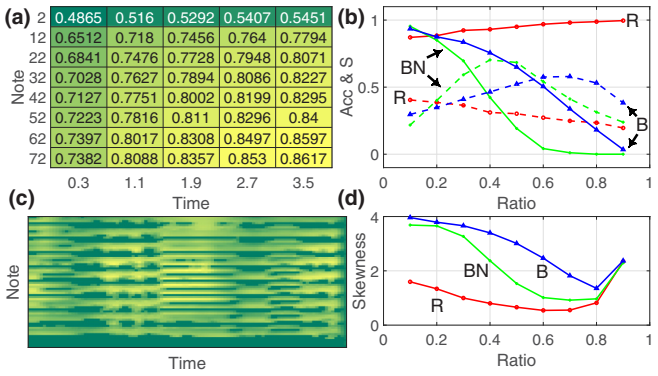

FIG. 7. (a) The accuracy of distinguishing the Baroque, Classical, and Romantic periods is shown with different sizes of 2D EEN information, as defined in the text. (b) Following the label in Fig. 3(a), the dotted and solid lines represent Shannon entropy (S) and accuracy (ACC). The features of the Romantic period are enhanced by discarding information. (c) The 2D EEN represents the features considered by Grad-CAM when back-propagating through the CNN to evaluate the period. The color indicates the scoring results of different regions, which also reflect the characteristics exhibited by groups of words. (d) Based on the magnitude of skewness, we can observe that the change in feature distribution due to the degradation of structure in the Romantic period is smaller than in the other two periods.

B. Minimum features and robustness of text networks

In addition to answering how many words are required to distinguish musical periods, we are ready to address question 4. Figure 6(a) summarizes the accuracy for different sizes of 2D EEN, which shows that, to obtain a test accuracy exceeding 80%, a minimum requirement is 42 notes/1.9 sec for task 1. As for task 2, we systematically increase the percentage of words removed from a text until a trained CNN changes its prediction. The Softmax layer in a CNN can convert features into probabilities for predicting labels [41], which allow us to compute Shannon entropy for these three periods. By analyzing 1000 samples from each period, we find in Fig. 7(b) that the accuracy of the Baroque and Classical samples decrease with more deletions. Surprisingly, the accuracy of the Romantic samples improves. This can be interpreted from hindsight as that the CPP music before the Romantic period emphasized a more rigorous pursuit of musical form and, therefore, allows less tolerance for arbitrary disruptions [37] in the coordination between different musical elements. In order to understand the characteristics of word distribution in a 2D EEN, we use grad CAM to extract the significant locations from the CNN [42]. Grad CAM, through the concept of back propagation, allows us to explore the scoring criteria for labeling [43]. Exemplifying a sample of 2D EEN, the color distribution in Fig. 7(c) ranging from green to yellow represents the scores of each word as calculated by Grad-CAM that functions as an indicator for the importance of each word when the CNN identifies the period label of the 2D EEN.

In Fig. 7(d) we use skewness, a good measure of distributional asymmetry, to calculate and plot the score histogram by averaging over the samples after Grad-CAM. We can observe that using skewness as a feature allows us to achieve temporal ordering between three periods based on the Grad-CAM scoring criteria. It is also evident that, in the context of random word disruption, Beethoven's representation of the Classical period tends to exhibit skewness closer to the Romantic period, while the decay trend in the Baroque period is comparatively gradual. The fact that the skewness is larger for the two periods before the Romantic implies the former did not have as many prominent features. This also means that the features were considered in a more holistic manner. Additionally, the score distribution during the Romantic period is closer to Gaussian. We can observe that, as the text structure gradually breaks down, indicating a departure from strict structural conventions, the less disrupted and more intact conditions closely resemble the Baroque period. Conversely, as the degree of disruption increases, the features tend to approach those of the Romantic period. The fact that this conclusion is reproduced in Fig. 7 vindicates the strength and accuracy of our approach.

C. Difference between music and nonmusic and the evolution of words

We attempt to address the difference between music and nonmusic through the fifth question. It is widely believed that the origin of music can be traced back to ancient humans imitating natural and ambient sounds [44,45]. To investigate this idea, we trained a network to distinguish between music and nonmusic including ambient sound and noises by 64

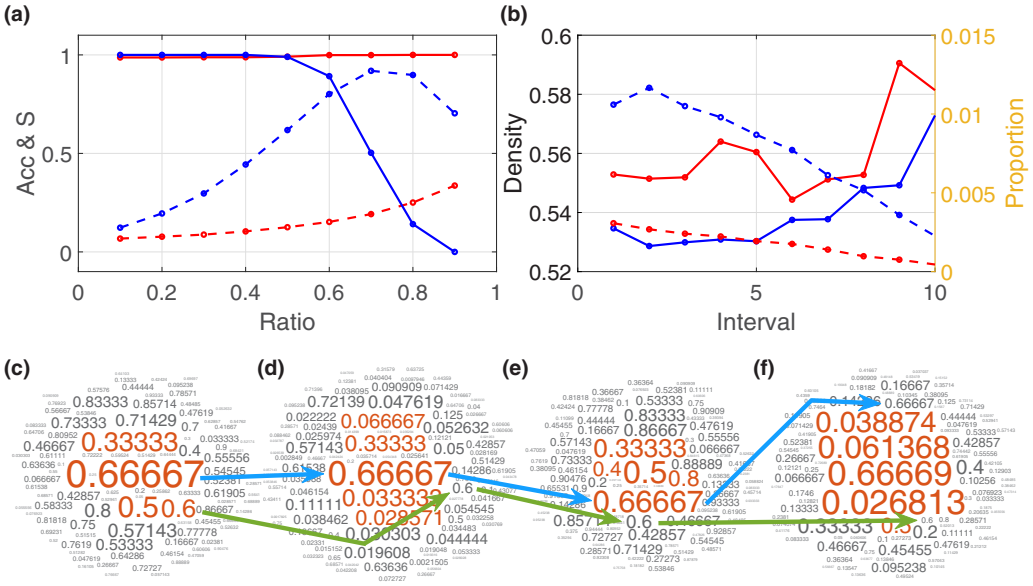

FIG. 8. In (a) and (b) the blue circle and red cross denote nonmusic and music, respectively. In (a) the solid and dashed lines represent accuracy (Acc) and Shannon entropy (S). (b) The dashed/solid line follows the note of the right/left y axes. Panels (c)–(f) show the term changes of Baroque, Classical, and Romantic periods and modern Japanese composers' music, such as those by Ryuichi Sakamoto, respectively. The destruction of terms is demonstrated by the termination of arrows, and new words can be seen to emerge.

notes and 10.1 sec. By the same process in Fig. 7(b), we detected the characteristics of 2D EEN in Fig. 8(a) and found that random destruction of the structure actually preserved its characteristics. We know that ambient sound and noise lack clear rhythm, melody, and harmony. However, their Shannon entropy reaches maximum when the loss rate exceeds 0.8. This implies that the word loss renders the ambient sound more music-like. We suspect that the deletion causes the originally continuous noise to become a combination of discrete segments, which is a feature of music. When the loss rate approaches 0.9, all nonmusic is transformed into music.

In Fig. 8(b) we normalized the Grad-CAM score into 10 equal parts, not only to calculate the proportion of 2D EEN but also to understand the grouping structure of Fig. 7(c). We used graph theory to calculate the density, $2L/[N(N-1)]$ where L, N denotes the link and node numbers. A link is defined if the distance is less than the average separation between each score point and all its neighbors. We found that although nonmusic has a higher proportion than music in each scoring part, its density is lower than that of music. This suggests that nonmusic characteristics have more scattered features than music. We also analyzed the word frequency of texts from the three periods of CPP and Japanese modern music using word cloud technology [46]. Since CC = 1 remains the most frequent word in all periods, we removed it to concentrate on the rest of the words in Figs. 8(c)–8(f). They show that words in EEN evolve throughout each period, while some are eventually terminated, just like words in natural language.

V. CONCLUSION AND DISCUSSIONS

By mapping the words of musical texts into one and two dimensions in EEN, we discovered different regularities in the composition structures of Baroque, Classical, and Romantic period music. Our approach from a more scientific view allows us to not only (1) obtain several results that are in line with current musical understandings, but also (2) differentiate nonmusic through its higher graph density from music, and offer insights into the evolution of musical texts. Among the conclusion for (1), we found that (i) the Baroque period is characterized by a more rigorous and ordered structure that emphasizes repeating the same form, as exemplified by the repetition of a particular pattern in works like fugues and Johann Pachelbel's Canon [47,48], (ii) although Beethoven in his position in the Classical period falls between the Baroque and Romantic periods, the characteristics of his music are closer to the structure of the Baroque period, and (iii) the arrangement of words for Romantic music distinctly differs from the two preceding periods, which supports the emphasis on individualism by Romantic composers.

Preliminary results for (2) suggest that it is promising to promote EEN to music with non-equal temperament laws. For instance, we may use the weights of Austronesian music as an indicator for the evolution and migration of Austronesian peoples [49,50], much like the role of genes [16]. Consequently, our approach has good potential to substantially advance the field of anthropology [51]. Another potential application is the study of animal languages. This is because EEN's ability to organize sound into textual systems can be extended to the ambient sound and Morse code as well. This means that it is plausible to analyze the sounds uttered by animals in various circumstances, especially in light of recent discoveries such as the apparent change in song patterns observed in white-throated sparrows across Canada [52]. The same concept can also offer some assistance to phonology. Our method explores how different phonetic elements are concatenated and combined to form a specific language. Within the framework of EEN, the weighted properties can offer discussions and classifications on the ways of linking and organizing these elements.

Comparing with previous methods that have claimed to find Zipf's law in music, our EEN focuses on the correlations within EEs, with an additional reference to the volume elements and weighting of EE contributions. One previous study used chords in musical notation to represent words without elaborating on the interword relationships [21]. In the meantime, another earlier work that emphasized psychoacoustics and encoded the usage of specific frequencies that are sensitive to the human auditory system also failed to consider the comprehensive adoption of *all* fundamental musical elements [20] in our view.

If the 2D EEN is the result of converting music into digitized musical text, it implies that if we can automatically generate a 2D EEN generator trained on a specific period, we could potentially use such a text-based network approach for music generation, for example, by utilizing the generated antagonism network (GAN) [53], which is different from directly manipulating music by audio format [54]. In the meantime, it is recommended to incorporate cycle-GAN [55] between 2D EEN and the Mel spectrum [56,57] to explore how words are presented in music information. In addition to the aforementioned sophisticated generative algorithms, one simpler numerical calculation can perhaps achieve the similar effects. It only requires the provision of (1) the distribution for Zipf's law containing *CC*s, (2) an adjacent matrix, (3) 2D EEN data, and (4) weights and thresholds. This information can then be utilized to adjust the volumes using Eq. (2). Based on the positive preliminary efforts, we are hopeful that the above processes can offer a straightforward approach to reconstruct volumes and achieve the music generator.

Due to the existence of 2D EEN, we can derive the probability distribution by statistically analyzing the relationships among various CC neighbors. The information contained in this distribution is then equivalent to defining the grammars. The above approach is akin to the method of random fields [58], which may help identify higher-level semantics beyond individual words and their treelike topology [59].

A knowledge graph, through processes such as numerical and vector representation, i.e., Word2VEC and Harmonic mean [60,61], aims to deduce the semantic relationships between textual concepts and their interconnections. By leveraging entity semantics, it forms a network of knowledge. Our text network EEN, denoted as CC, can serve as a scalar for textual information. Furthermore, the text network inherently organizes contextual relationships to help us understand the semantic meaning behind music. For the increasingly active field of natural language processing in deep learning, there is a widespread lack of interpretability. Remedying this problem, our definition of words can directly correspond to the actual power spectrum.

Schopenhauer thinks that music is an embodiment of will [62], and how exactly emotions are expressed through music has always been a topic of debate [63]. By offering a basis to extract the meaning behind music in a systematic and quantitative way, we show that there may indeed be a "language of the emotions" [63] that musicians have cultivated throughout history.

ACKNOWLEDGMENT

Financial support from the National Science and Technology Council in Taiwan under Grants No. 111-2112-M007-025 and No. 112-2112-M007-015 is acknowledged.

[1] J. N. Chiang, M. H. Rosenberg, C. A. Bufford, D. Stephens, A. Lysy, and M. M. Monti, The language of music: Common neural codes for structured sequences in music and natural language, Brain Language **185**, 30 (2018).

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

[3] J. McDermott, The evolution of music, Nature (London) **453**, 287 (2008).

[4] K. N. Smith, Languages are products of their environments, *Discover Magazine* (2015).

[5] I. Maddieson and C. Coupé, Human spoken language diversity and the acoustic adaptation hypothesis, J. Acoust. Soc. Am. **138**, 1838 (2015).

[6] D. Douglass, Issues in the use of IA Richards' tenor-vehicle model of metaphor, Western J. Commun. **64**, 405 (2000).

[7] I. A. Richards and J. Constable, *The Philosophy of Rhetoric* (Routledge, London, 2018).

[8] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. D. Friederici, Music, language and meaning: Brain signatures of semantic processing, Nat. Neurosci. **7**, 302 (2004).

[9] B. Maess, S. Koelsch, T. C. Gunter, and A. D. Friederici, Musical syntax is processed in Broca's area: An MEG study, Nat. Neurosci. **4**, 540 (2001).

[10] S. Brown, A joint prosodic origin of language and music, Front. Psychol. **8**, 1894 (2017).

[11] M. Tettamanti and D. Weniger, Broca's area: A supramodal hierarchical processor? Cortex **42**, 491 (2006).

[12] S. Koelsch, Toward a neural basis of music perception—A review and updated model, Front. Psychol. **2**, 110 (2011).

[13] M. Maiden, Irregularity as a determinant of morphological change, J. Linguist. **28**, 285 (1992).

[14] P. Durrant, Formulaicity in an agglutinating language: The case of Turkish, Corpus Linguistics Linguistic Theory **9**, 1 (2013).

[15] S. T. Piantadosi, Zipf's word frequency law in natural language: A critical review and future directions, Psychon. Bull. Rev. **21**, 1112 (2014).

[16] C. Furusawa and K. Kaneko, Zipf's law in gene expression, Phys. Rev. Lett. **90**, 088102 (2003).

[17] A. Broder, R. Kumar, F. Maghoul, P. Raghavean, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Graph structure in the web, Comput. Netw. **33**, 309 (2000).

[18] P.-R. Tsai, K.-H. Chen, T.-M. Hong, F.-N. Wang, and T.-Y. Huang, Categorizing SHR and WKY rats by chi2 algorithm and decision tree, Sci. Rep. **11**, 3463 (2021).

[19] V. M. Eguíluz, D. R. Chialvo, G. A. Cecchi, M. Baliki, and A. V. Apkarian, Scale-free brain functional networks, Phys. Rev. Lett. **94**, 018102 (2005).

[20] M. Haro, J. Serrá, P. Herrera, and Á. Corral, Zipf's law in short-time timbral codings of speech, music, and environmental sound signals, PLoS ONE **7**, e33993 (2012).
[21] J. I. Perotti and O. V. Billoni, On the emergence of Zipf's law in music, Physica A **549**, 124309 (2020).
[22] D. Teie, A comparative analysis of the universal elements of music and the fetal environment, Front. Psychol. **07**, 1158 (2016).
[23] P. BehnamGhader *et al.*, LLM2Vec: Large language models are secretly powerful text encoders, arXiv:2404.05961.
[24] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).
[25] P. N. Juslin and R. Timmers, *Handbook of Music and Emotion: Theory, Research, Applications* (Oxford University Press, Oxford, 1993), p. 452.
[26] S. Hallam, Musical motivation: Towards a model synthesizing the research, Music Ed. Res. **4**, 225 (2002).
[27] G. Karl, Structuralism and musical plot, Music Theory Spectrum **19**, 13 (1997).
[28] V. Sorin, Y. Barash, E. Konen, and E. Klang, Deep learning for natural language processing in radiology—Fundamentals and a systematic review, J. Am. Coll. Radiol. **17**, 639 (2020).
[29] S. Kuang and B. D. Davison, Class-specific word embedding through linear compositionality, in *Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (IEEE, Piscataway, NJ, 2018), p. 1.
[30] Z. Sadeghi, J. L. McClelland, and P. Hoffman, You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes, Neuropsychologia **76**, 52 (2015).
[31] K. Tedman and E. Varése, Concepts of organized sound, D.Phil. dissertation, University of Sussex, 1983.
[32] D. Tymoczko, *Geometry of Music Harmony and Counterpoint in the Extended Common Practice* (Oxford University Press, Oxford, 2014).
[33] R. P. Morgan, Symmetrical form and common-practice tonality, Music Theory Spectrum **20**, 1 (1998).
[34] J. Harbison, Symmetries and the "new tonality", Contemp. Music Rev. **6**, 71 (1992).
[35] H. Cho and S. M. Yoon, Issues in visualizing intercultural dialogue using Word2Vec and t-SNE, in *Proceedings of the 2017 International Conference on Culture and Computing (Culture and Computing)* (IEEE, Piscataway, NJ, 2017), pp. 149–150.
[36] G. S. Johnston, Polyphonic keyboard accompaniment in the early Baroque: An alternative to basso continuo, Early Music **26**, 51 (1998).
[37] V. K. Agawu, *Music as Discourse: Semiotic Adventures in Romantic Music* (Oxford University Press, Oxford, 2014).
[38] C. Dahlhaus and E. Sanders, Review: Romantic music: A history of musical style in nineteenth-century Europe, 19th Century Music **11**, 194 (1987).
[39] N. Milosevic, *Introduction to Convolutional Neural Networks: With Image Classification Using PyTorch* (Apress, New York City, 2020).
[40] R. Albert, H. Jeong, and A.-L. Barabási, Error and attack tolerance of complex networks, Nature (London) **406**, 378 (2000).
[41] R. Hu, B. Tian, S. Yin, and S. Wei, Efficient hardware architecture of softmax layer in deep neural network, in *Proceedings of the IEEE 23rd International Conference on Digital Signal Processing (DSP)* (IEEE, Piscataway, NJ, 2018).
[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vision **128**, 336 (2020).
[43] K. Vinogradova, A. Dibrov, and G. Myers, Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract), Proc. AAAI Conf. Artif. Intell. **34**, 13943 (2020).
[44] W. T. Fitch, The biology and evolution of music: A comparative perspective, Cognition **100**, 173 (2006).
[45] S. Mithen, *The singing Neanderthals: The Origins of Music, Language, Mind, and Body* (Harvard University Press, Cambridge, MA, 2005).
[46] W. Cui, Y. Wu, S. Liu, F. Wei, M. Zhou, and H. Qu, Context preserving dynamic word cloud visualization, in *Proceedings of the 2010 IEEE Pacific Visualization Symposium (PacificVis)* (IEEE, Piscataway, NJ, 2010).
[47] S. Gut, Review: Music analysis in the nineteenth century. Vol. 1: Fugue, form and style, Revue Musicol. **85**, 373 (1999).
[48] C. Agon and M. Andreatta, Modeling and implementing tiling rhythmic canons in the openmusic visual programming language, Perspect. New Music **49**, 66 (2011).
[49] T. Rzeszutek, P. E. Savage, and S. Brown, The structure of cross-cultural musical diversity, Proc. R. Soc. B **279**, 1606 (2011).
[50] B. Abels, *Austronesian Soundscapes Performing Arts in Oceania and Southeast Asia* (Amsterdam University Press, Amsterdam, 2011).
[51] S. Brown *et al.*, Correlations in the population structure of music, genes and language, Proc. R. Soc. B **281**, 20132072 (2014).
[52] K. A. Otter *et al.*, Continent-wide shifts in song dialects of white-throated sparrows, Curr. Biol. **30**, 3231 (2020).
[53] X. Mao and Q. Li, *Generative Adversarial Networks for Image Generation* (Springer, Berlin, 2020).
[54] N. Jiang, S. Jin, Z. Duan, and C. Zhang, Rl-duet: Online music accompaniment generation using deep reinforcement learning, Proc. AAAI Conf. Artif. Intell. **34**, 710 (2020).
[55] J. Harms, Y. Lei, T. Wang, R. Zhang, J. Zhou, X. Tang, W. J. Curran, T. Liu, and X. Yang, Paired cycle-GAN-based image correction for quantitative cone-beam computed tomography, Med. Phys. **46**, 3998 (2019).
[56] Y. Khasgiwala and J. Tailor, in *Proceedings of the 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON 2021)—Report, Power and Communication Technologies (GUCON)* (IEEE, Piscataway, NJ, 2021), p. 339.
[57] N. Perraudin, P. Balazs, and P. L. Sondergaard, A fast Griffin-Lim algorithm, in *Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE, New York, 2013).
[58] M. Brett, W. Penny, and S. Kiebel, Introduction to random field theory, in *Human Brain Function*, 2nd ed., edited by R. Frackowiak, K. Friston, C. Fritch, R. Dolan, C. Price, and W. Penny (Academic Press, Burlington, 2004), pp. 867–880.
[59] S. Dehaene *et al.*, The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees, Neuron **88**, 2 (2015).
[60] E. Palumbo, G. Rizzo, R. Troncy, E. Baralis, M. Osella, and E. Ferro, Knowledge graph embeddings with node2vec for item

recommendation, in *The Semantic Web: ESWC 2018 Satellite Events, Heraklion, Crete, Greece, Revised Selected Papers 15* (Springer International Publishing, Cham, 2018).

[61] F. Akrami, M. S. Saeef, Q. Zhang, W. Hu, and C. Li, Realistic re-evaluation of knowledge graph completion methods: An experimental study, in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (ACM Press, New York, 2020), pp. 1995–2010.

[62] A. Schopenhauer, *Die Welt als Wille und Vorstellung* (BoD—Books on Demand, Norderstedt, Germany, 2016).

[63] A. Kania, The philosophy of music, in *The Stanford Encyclopedia of Philosophy*, Spring Edition, edited by E. N. Zalta and U. Nodelman (Stanford University Press, Stanford, CA, 2023).