HIERARCHICAL GRAPH-BASED NEURAL NETWORK FOR SINGING MELODY EXTRACTION

Shuai Yu [1], *Xi Chen* [1] *and Wei Li** [1,2]

[1] School of Computer Science and Technology, Fudan University, Shanghai, China
[2] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China

ABSTRACT

Singing melody extraction from polyphonic music is a critical and challenging task in music information retrieval (MIR). However, due to the interfere of the accompaniment and the background noise, it is key and challenging to obtain a global semantic representation that discriminates the singing melody line. To address this issue, we consider the two aspects that regards to obtaining the global semantic representation: the global relationships in the spectrum and the relationships between channels. In this paper, we propose a novel hierarchical graph-based network for singing melody extraction. In particular, according to its characteristics of the spectrum, we first model the spectrum into graph structure, a two-layer graph convolution network is used to obtain the global semantic representation in the spectrum. Then to capture the relationships between channels, channel-wise graph convolution module is devised to capture and reasoning the relationship between channels. The conducted experiments demonstrate the effectiveness of the proposed network.

Index Terms— Singing Melody Extraction, Graph-based Neural Network, Music Information Retrieval

1. INTRODUCTION

Singing melody extraction is a challenging task in music information retrieval (MIR), which estimates the fundamental frequency (F0) of the leading voice. It has many downstream applications, such as music recommendation [1] and cover song identification [2] in stream media platforms. The difficulty of this task is the presence of background sounds interwoven with the leading voice. The extraction in a noisy environment has already been a difficult problem [3]. It gets even more difficult to extract the leading voice when the background sounds are musical accompaniments. Elements in accompaniments like chord progression will naturally contain the leading vocal F0 or its harmonics. As a consequence, it is not trivial to obtain semantic representation that discriminates the singing melody from the background music.

Recently, a large number of approaches are proposed to learn better semantic representation for singing melody extraction [4, 5, 6, 7, 8]. Researchers have employed standard convolutional layers to learn local patterns and use pooling layers to select the most prominent features. With stacked operations, the receptive filed of CNNs can be able to access larger regions of the input spectrum and eventually obtain a semantic feature for singing melody extraction. Hsieh et al. [5] employed a streamlined architecture to learn non-linear projection between audio and labels for singing melody extraction. However, due to the fixed structures, such fashion are limited to local receptive field, which will lead to octave errors (several octaves above the actual pitch of the singer) and other tone errors. In addition to the standard convolution fashion, researchers also attempted to use dilated convolution neural network to enlarge the receptive field of CNNs for obtaining semantic representation. Gao et al. [6] employed a set of dilated convolutions with different dilated rates to fuse feature maps from different levels. Unfortunately, the dilated convolution solution can only learn information from surrounding pixels which can hardly learn global information.

Fig. 1: Diagram (a) is ground truth. Diagram (b) and (c) are two channels that extracted from pretrained MSNet [5]. Diagram (b) contains the melody line and its harmonics. Diagram (c) contains the melody line and its sub-harmonics. Diagram (d) schematically expresses relationships between channels.

Although spatial attention mechanism [9] is not applied to this task yet, it has been proven effective on computer vision tasks [10, 11]. The spatial attention mechanism learns non-local relationship in a feature map which can obtain a better semantic representation. However, spatial attention mechanism is computationally expensive. Different from melody extraction task, the inputs of singing melody extraction are 2-D high-resolution spectrums, the space complexity is $O((F \times T) \times ((F \times T))$, where $F \times T$ denotes the frequency and time dimension of the input feature map. We argue that spatial attention mechanism is not practical for this task. To design a light-weight model, in this paper, we adopt a graph-based network to capture global and local information in the spectrum for singing melody extraction.

Through observing the channels in the CNN based networks for melody extraction, we found that there are certain correlations between feature maps in different channels. As is shown in Fig. 1, some feature maps contain the melody line (F0) and its harmonics, while some feature maps contain the melody line and its sub-harmonics. We assume that relationships between channels are necessary to be captured and learned. However, existing channel-wise processing techniques, such as the channel attention mechanism [12, 11], generally view each feature map as a whole. The contour details in the channel are neglected when learning interaction between channels. Therefore a finer-grained network is needed to learn such relationships. To this end, we devise a novel channel-wise graph convolution block to learn the relationship between channels.

In summary, the contribution of this paper is twofold: i) we propose a hierarchical graph-based network for singing melody ex-

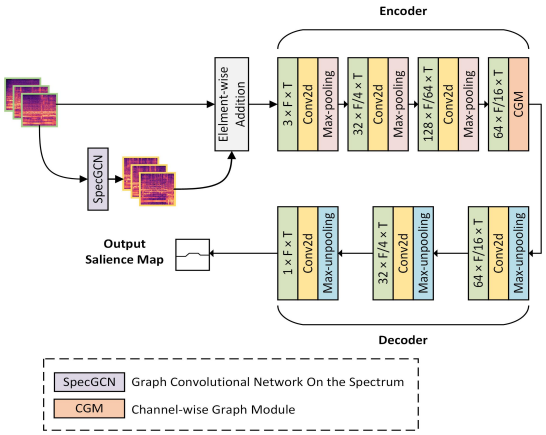


Fig. 2: The overall architecture of our proposed method.

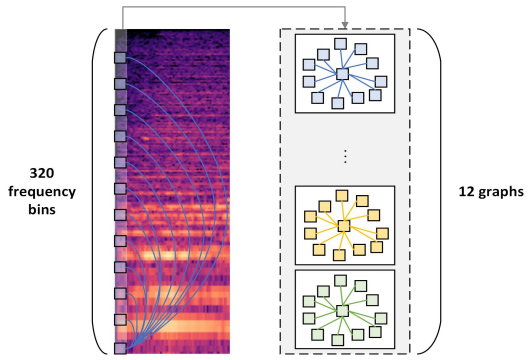


Fig. 3: Illustration of the procedures of modeling the spectrum into graph structure.

traction to learn global and local representation in the spectrum and channels. Specifically, the graph convolution layers in the spectrum learns the global semantic representation in the spectrum. The graph convolution layers between channels are responsible for reasoning the relationship between channels. ii) We use part of the vocal tracks of the MedleyDB dataset for training the model and we evaluate the performance on ADC2004, MIREX 05 and MedleyDB (no overlapping) datasets. The experimental study demonstrates the effectiveness of our method compared with other state-of-the-art ones.

2. OUR METHODS

The overall architecture of our proposed hierarchical graph-based network is illustrated in Fig. 2. In this section, we will firstly introduce the input of the model and then develop our model in detail.

2.1. Input Representation

In this paper, we choose to use a set of mid-level representations as the input of our model. It contains three parts: (1) the generalized cepstrum(GC) [13], (2) the generalized cepstrum of spectrum (GCoS) [14], (3) the Combined Frequency and Periodicity (CFP) spectrum [15]. In this work, the audio files are resampled to 8 kHz and merged into one mono channel. Data representations are computed with a Hanning window of 768 samples and hop size of 80 samples.

Given the magnitude of the STFT of an input signal $\mathbf{X}$, the CFP can be calculated as follow:

$$\mathbf{Z}_S = \sigma_0(\mathbf{W}_f\mathbf{X}), \tag{1}$$

$$\mathbf{Z}_{GC} = \sigma_1(\mathbf{W}_t\mathbf{F}^{-1}\mathbf{Z}_\mathbf{S}), \tag{2}$$

$$\mathbf{Z}_{GCoS} = \sigma_2(\mathbf{W}_f\mathbf{F}\mathbf{Z}_{GC}), \tag{3}$$

where $\mathbf{W}_f$ and $\mathbf{W}_t$ are used to remove slow-varying portions. $\mathbf{F}$ denotes a DFT matrix and σ_i is activation function [16]. Two sets of filter banks are applied in the time and frequency domains, respectively. Due to the page limitation, readers can refer to [13, 14] for more details. The final CFP representation is calculated as follow:

$$\mathbf{Z}_{CFP} = \hat{\mathbf{Z}}_{GC} \times \hat{\mathbf{Z}}_{GCoS} \tag{4}$$

where $\hat{\mathbf{Z}}_{GC}$ and $\hat{\mathbf{Z}}_{GCoS}$ are filtered representations. To be clear, the input of our model is:

$$I = [\mathbf{Z}_{GC}, \mathbf{Z}_{GCoS}, \mathbf{Z}_{CFP}] \tag{5}$$

2.2. Hierarchical Graph-based Network

In this work, we employ MSNet [5] as the backbone of the proposed model due to its effectiveness and robustness. And we hope that the performance gains can be easily attributed to through such settings.

2.2.1. Graph Convolution Network In the Spectrum (SpecGCN)

It is well known that graph convolution network is initially designed for modeling non-Euclidean data. However, the input spectrum belongs to Euclidean data. Therefore, we need first introduce how we model the Euclidean-based spectrum into a graph structure and then feed it into graph convolution networks. The procedures of modeling the spectrum is illustrated in Fig. 3.

Previous works for singing melody extraction often treated the spectrum as an image, so that advanced techniques in computer vision can be applied to this task. However, the spectrum has its inherent characteristics which ordinary images do not have. The most important characteristic of the spectrum is harmonic relationship along the frequency axis and temporal relationship along the time axis. Motivated by such characteristics, we take 5×64 pixels as a node in the graph, which denotes 5 frequency bins (corresponding to one semitone) and 64 time steps (corresponding to 640 ms). By such setting, the spectrum now can be grouped into many nodes with dimension of 320 (5×64), the nodes are then connected with the nodes at the harmonic multiples.

To be self-contained, we briefly introduce the graph convolution network to capture global relationship in the spectrum. Given graph $G = (V, E)$ and its adjacency matrix A and degree matrix D, the normalized graph Laplacian matrix L is defined as:

$$L = I - D^{-1/2}AD^{-1/2} \tag{6}$$

With a renormalizing trick [17], the propagation rule of a graph convolution network is given:

$$H^{l+1} = \sigma(\widetilde{D}^{-1/2}A\widetilde{D}^{-1/2}H^l\Theta^l) \tag{7}$$

where H^l is the node features of l-th layer, Θ is parameters to be trained at layer l and σ is the non-linear activation function.

After obtaining the propagation rule of graph convolution network, we use a two-layer graph convolution network on our designed graph structure $\mathbf{S}$ in the spectrum:

$$\mathbf{S}' = \phi'(GCN(\phi(GCN(\mathbf{S})))) \tag{8}$$

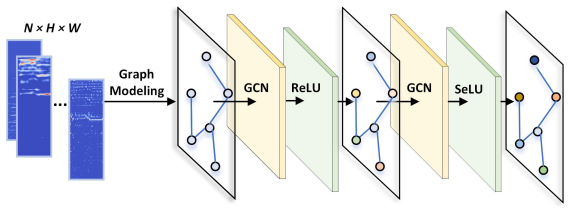


Fig. 4: Illustration of the channel-wise graph convolution module.

where $\mathbf{S}'$ has the same dimensions with $\mathbf{S}$ and GCN denotes the graph convolution operation. ϕ and ϕ' are the *ReLu* and *Normalize* functions. We pass the message of each node in the graph by graph convolution and the harmonic relationship can be captured and learned. Note that, the edges between nodes are unidirectional. We assume such operation will help to obtain global semantic representation for singing melody extraction. After we obtaining S', we perform an element-wise addition on S' and the original inputs for thereafter training.

2.2.2. *Channel-wise Graph Convolution Module (CGM)*

We notice that there are certain correlations between feature maps in different channels. However, previous works treated the channel as a whole (e.g., using global average pooling to transform a feature map into a number) and only learn the weights on the channel-level which cannot really learn the fine-grained relationships between channels as mentioned in the Sec. 1.

To this end, we devise a novel channel-wise graph convolution performed on the channels. As Fig. 4 is shown, two stacked GCN layers are performed on the graph structure. Formally, given a set of channels $\mathbf{C} \in \mathbb{R}^{N \times H \times W}$, where N is the number of the channels. For each channel c, we first reshape it to $c \in \mathbb{R}^{HW}$, then a non-linear transformation is performed to project the feature map c into same hidden space:

$$\mathbf{c}' = Dropout(ReLU(GCN(\mathbf{c}))) \tag{9}$$

where $ReLU$ is the non-linear activation function, c' is the transformed feature map. To avoid overfitting, we use an additional dropout layer to achieve it. To improve the capacity of our proposed network, another graph convolution layer is applied on the transformed feature map $\mathbf{c}'$.

To combine with the standard convolution operation, we need to reverse the projection matrix into the origin shape:

$$c'' = SeLU(GCN(\mathbf{c}')) \tag{10}$$

where $c'' \in \mathbb{R}^{H \times W}$. Note that, the edges between nodes are unidirectional. In the experiment, we found that using bidirectional edges will decrease the performance. By including CGM, we believe that the model can learn the finer-grained relationship between channels and can achieve better performance on this task.

3. EXPERIMENT

3.1. Experiment Setup

We randomly choose 60 vocal tracks from the Medley DB dataset. To increase the amount of the training data, we augment the training dataset by copying some of the chosen vocal tracks. Accordingly, there are 98 clips in the training dataset. We select only samples having melody sung by human voice from ADC2004, MIREX 05[1] and MedleyDB for test sets. As a result, 12 clips in ADC2004, 9 clips in MIREX 05 and 12 clips in MedleyDB are selected. Note that there is no overlap between the training and testing datasets.

Method	ADC 2004 (vocal)		
	OA	RPA	RCA
Backbone	81.4	82.7	84.9
Backbone + SpecGCN	82.1	83.8	84.1
Backbone + CGM	83.1	83.5	84.2
Backbone + SpecGCN + CGM	**83.9**	**84.8**	**85.3**
Backbone + Channel Attn.	81.7	83.5	85.4
Backbone + Spatial Attn.	80.2	81.5	82.2

Table 1: Ablation study results of the proposed model on the ADC 2004 dataset. The values in the table are percentile.

3.2. Implementation Details

For the hyperparameters in computing CFP, the number of frequency bins is set to 320, with 60 bins per octave, and the frequency range is from 31 Hz (B0) to 1250 Hz (D#6). We use MSNet [5] as our backbone due to its effectiveness and popularity. When graph convolution is applied on the modeled graph structure in the spectrum, the hidden dimensions of the graph convolution layer are 120. For parameters of CGM modules, the number of GCN layers is 2 and the hidden dimensions are 64. Here, we use dropout mechanism to avoid overfitting. The dropout rate is 40%. We train our model for 100 epoches in total, with an inital learning rate 0.0002. Adam is employed as the optimizer of our network. The model is trained with binary cross entropy loss function.

3.3. Evaluation Metrics

Following the convention in the literature [18], we use the following metrics for performance evaluation: overall accuracy (OA), raw pitch accuracy (RPA), raw chroma accuracy (RCA), voicing recall (VR) and voicing false alarm (VFA). OA is often considered more important among these metrics.

3.4. Baseline Methods

We compare the proposed methods with three recently developed state-of-the-art deep learning algorithms: The first is MD+MR proposed by [6], which employed a set of dilation convolutions for this task. The second one is patch-based CNN proposed by [15], which is a state-of-the-art method for singing melody extraction. The third one is MSNet proposed by [5], which is a very strong baseline. Our model is trained and tested on a machine with NVIDIA GTX 1080 Ti GPU and 32GB RAM. Our model is implemented with PyTorch [2].

3.5. Ablation Study

To verify the effectiveness of our proposed hierarchical graph-based network, we design a set of ablation studies to justify the performances in different cases. We first evaluate the performance of the proposed hierarchical graph network. If we remove the hierarchical graph network, the model remains backbone. Then we respectively

[1] https://labrosa.ee.columbia.edu/projects/melody

[2] https://www.pytorch.org

Method	ADC 2004 (vocal)				
	OA	RPA	RCA	VR	VFA
MD+MR [6]	82.8	82.4	84.6	85.3	**12.7**
PATCH CNN [15]	74.3	76.7	78.4	**90.1**	41.3
MSNet [5]	81.4	82.7	84.9	87.4	18.6
Proposed	**83.9**	**84.8**	**86.1**	89.2	21.3

(a) ADC 2004

Method	MIREX 2005 (vocal)				
	OA	RPA	RCA	VR	VFA
MD+MR [6]	80.7	78.6	79.5	84.3	**16.9**
PATCH CNN [15]	74.4	83.1	83.5	**95.1**	41.1
MSNet [5]	78.6	78.4	79.7	85.7	21.5
Proposed	**81.3**	**85.2**	**86.4**	93.2	21.7

(b) MIREX 05

Method	Medley DB (vocal)				
	OA	RPA	RCA	VR	VFA
MD+MR [6]	62.3	52.2	57.4	62.9	21.5
PATCH CNN [15]	61.5	62.8	63.4	64.2	27.6
MSNet [5]	65.8	59.3	64.8	**72.1**	27.9
Proposed	**67.9**	**61.2**	**65.8**	71.7	21.6

(c) MedleyDB

Table 2: Results of the proposed and baseline methods on the ADC2004, MIREX 05 and MedleyDB datasets. The values in the table are percentile.

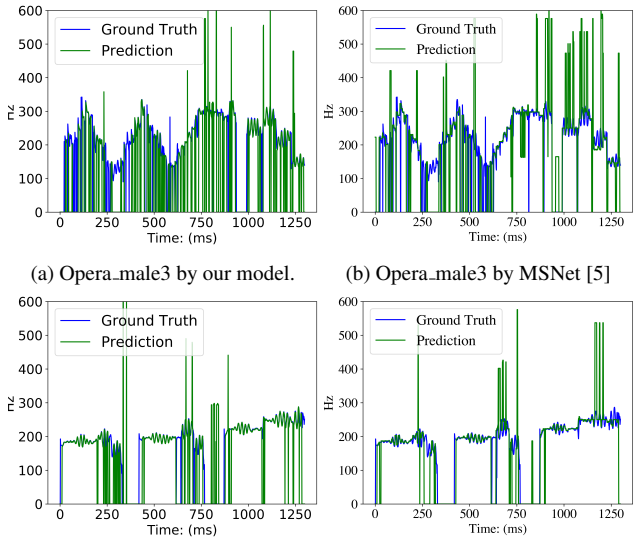
remove SpecGCN and CGM modules to evaluate the effectiveness of them. The performance is shown in Table 1, when focusing on OA, "Backbone + SpecGCN" outperforms "Backbone" by 0.8%. "Backbone + CGM" outperforms "Backbone" by 2.1%. Using both SpecGCN and CGM can achieve the best performance, when focusing on OA, "Backbone + SpecGCN + CGM" outperforms "Backbone" by 3.1%.

Since spatial attention can also learn the global semantic relationship in the spectrum, therefore it is necessary to compare the performance with our proposed SpecGCN. And though channel attention cannot really learn the finer-grained relationship between channels, it would be interesting to see how much we bridge the semantic gaps between the channel attention and proposed CGM modules.

As Table 1 is shown, the performance of the "Backbone + Spatial Attn" is slightly decreased than using Backbone only. The spatial attention captures relationships of all the pixels in the spectrum, however, the spectrum not only includes the harmonic relationship and contains other components such as accompaniment and background noise, we guess such operation may bring performance decrease. For channel attention, "Backbone + Channel Attn." slightly outperforms "Backbone". However, our proposed "Backbone + CGM" can significantly improve the "Backbone" than "Backbone+Channel Attn." .

3.6. Comparison with state-of-the-art methods

In this section, the proposed method is compared with three state-of-the-art baseline methods. Table 2 shows the results of the proposed and baseline methods on ADC2004, MIREX 05 and MedleyDB. The proposed model and the four baseline methods are trained on the same dataset. As shown in Table 2, the results show that the proposed method consistently outperforms other baseline methods in general. When focusing on OA, the proposed model outperforms MSNet by 3.07% on ADC2004, by 3.32% on MIREX05, and by 3.19% on Medley DB. We can also find that the proposed model achieved comparable VR and VFA. Since this paper focuses on melody extraction other than melody detection, we leave the improvement of VR and VFA as a future research topic. Table 2 also shows the proposed model consistently outperforms the baselines across the three datasets and validates the robustness of our model.



(a) Opera_male3 by our model. (b) Opera_male3 by MSNet [5]

(c) Opera_male5 by our model. (d) Opera_male5 by MSNet [5]

Fig. 5: Visualization of singing melody extraction results on two opera songs using different models.

To further evaluate the effectiveness of the proposed model, we visualize the predictions of the two opera singing songs on the ADC2004 dataset. As depicted in Fig. 5, we can observe that there are fewer octave errors in diagram (a) and (c) than in diagram (b) and (d). Through the visualization of the predicted melody contour, it indicates that the performance gains of the proposed model can be attributed to solving the octave errors and other tone errors.

4. CONCLUSION

In this paper, we have proposed hierarchical graph-based network for singing melody extraction. To enhance the representation, we proposed a hierarchical graph-based network to extract the relationship in the spectrum and the nodes between channels. We used part of the vocal tracks of MedleyDB dataset for training the model and evaluated the performance on the ADC 2004, MIREX 05 and MedleyDB. The experimental study demonstrates the superiority of our method compared with other state-of-the-art ones.

5. ACKNOWLEDGEMENT

This work was supported by National Key R&D Program of China (2019YFC1711800), NSFC(62171138).

6. REFERENCES

[1] Peter Knees and Markus Schedl, "Music retrieval and recommendation: A tutorial overview," in *Proc. SIGIR*, 2015, pp. 1133–1136.

[2] Joan Serra, Emilia Gómez, and Perfecto Herrera, "Audio cover song identification and similarity: background, approaches, evaluation, and beyond," in *Advances in Music Information Retrieval*, pp. 307–332. Springer, 2010.

[3] Joseph Tabrikian, Shlomo Dubnov, and Yulya Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *TASLP*, vol. 12, no. 1, pp. 76–87, 2004.

[4] Shuai Yu, Yi Yu, Xi Chen, and Wei Li, "HANME: hierarchical attention network for singing melody extraction," *IEEE Signal Process. Lett.*, vol. 28, pp. 1006–1010, 2021.

[5] Tsung-Han Hsieh, Li Su, and Yi-Hsuan Yang, "A streamlined encoder/decoder architecture for melody extraction," in *Proc. ICASSP*, 2019, pp. 156–160.

[6] Ping Gao, Cheng-You You, and Tai-Shih Chi, "A multi-dilation and multi-resolution fully convolutional network for singing melody extraction," in *Proc. ICASSP*, 2020, pp. 551–555.

[7] Shuai Yu, Xiaoheng Sun, Yi Yu, and Wei Li, "Frequency-temporal attention network for singing melody extraction," in *Proc. ICASSP*, 2021, pp. 251–255.

[8] Ke Chen, Shuai Yu, Cheng-i Wang, Wei Li, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Tonet: Tone-octave network for singing melody extraction from polyphonic music," *Proc. ICASSP*, 2022.

[9] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. CVPR*, 2017, pp. 5659–5667.

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018, pp. 3–19.

[11] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proc. CVPR*, 2017, pp. 3156–3164.

[12] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[13] Takao Kobayashi and Satoshi Imai, "Spectral analysis using generalised cepstrum," *TASLP*, vol. 32, no. 6, pp. 1235–1238, 1984.

[14] Li Su, "Between homomorphic signal processing and deep neural networks: Constructing deep algorithms for polyphonic music transcription," in *Proc. APSIPA ASC*, 2017, pp. 884–891.

[15] Li Su, "Vocal melody extraction using patch-based cnn," in *Proc. ICASSP*, 2018, pp. 371–375.

[16] Li Su and Yi-Hsuan Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *TASLP*, vol. 23, no. 10, pp. 1600–1612, 2015.

[17] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017.

[18] Justin Salamon, Emilia Gómez, Daniel PW Ellis, and Gaël Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.