한국음향학회지 제44권 제5호 pp. 516~523 (2025)
The Journal of the Acoustical Society of Korea Vol.44, No.5 (2025)
https://doi.org/10.7776/ASK.2025.44.5.516

pISSN : 1225-4428
eISSN : 2287-3775

MRNet: A multi-route convolutional neural network for robust music representation learning

음악 표현 학습을 위한 다중 경로 합성곱 신경망

Jungwoo Heo,[1] Hyun-seo Shin,[1] Chan-yeong Lim,[1] Kyo-won Koo,[2] Seung-bin Kim,[2]
Jisoo Son,[1] and Ha-Jin Yu[1†]

(허정우,[1] 신현서,[1] 임찬영,[1] 구교원,[2] 김승빈,[2] 손지수,[1] 유하진[1†])

[1]Department of Computer Science, University of Seoul
[2]Department of Artificial Intelligence, University of Seoul

(Received August 6, 2025; revised August 29, 2025; accepted August 30, 2025)

ABSTRACT: Music Information Retrieval (MIR) focuses on extracting semantic information embedded in audio signals, such as genre, artist identity, and tempo. These musical cues cover a wide range of temporal characteristics, from short-term features like pitch and timbre to long-term patterns such as melody and mood, and they require processing at multiple levels of abstraction. In this paper, we propose a Multi-Route Neural Network (MRNet) designed to capture musical representations that reflect both short-term and long-term characteristics, as well as different levels of abstraction. To achieve this, MRNet stacks several convolutional layers with different dilation rates, allowing the model to analyze audio patterns over multiple time scales. Additionally, it introduces a specialized module called the multi-route Res2Block, which separates the processing path into multiple branches. Each branch processes the input to a different depth, enabling the network to extract low-level, mid-level, and high-level features simultaneously. MRNet achieves classification accuracies of 94.5 %, 56.6 %, 63.2 %, and 71.3 % on the GTZAN, FMA Small, FMA Large, and Melon datasets, respectively, outperforming previous Convolution Neural Network(CNN)-based approaches. These results demonstrate the effectiveness of MRNet in learning robust and hierarchical music representations for MIR tasks.
Keywords: Deep learning, Music information retrieval, Convolution Neural Network (CNN), Music representation

PACS numbers: 43.75.Zz, 43.60.Lq

초 록: 음악 정보 검색(Music Information Retrieval, MIR)은 오디오 신호에 내재된 장르, 아티스트 정체성, 템포와 같은 의미적 정보를 추출하는 데 중점을 둔 연구 분야이다. 이러한 음악적 단서들은 피치나 음색과 같은 단기적 특성부터 멜로디나 분위기와 같은 장기적 패턴에 이르기까지 다양한 시간적 특성을 포함하며, 여러 수준의 추상화된 처리를 요구한다. 본 논문에서는 음악의 단기적 특성과 장기적 특성은 물론, 다양한 추상화 수준을 모두 반영할 수 있는 음악 표현을 학습하기 위해 설계된 다중 경로 신경망(Multi-Route Neural Network, MRNet)을 제안한다. 이를 위해 MRNet은 수용 영역의 크기가 서로 다른 여러 개의 확장 합성곱 계층을 적층하여, 다양한 시간 범위에 걸친 오디오 패턴을 효과적으로 분석할 수 있도록 한다. 또한, 입력 신호를 여러 경로로 나누어 처리하는 특수 구조인 multi-route Res2Block을 도입하여, 각 경로에서 서로 다른 깊이로 특징을 추출할 수 있게 설계하였다. 이 구조를 통해 네트워크는 저차, 중차, 고차 수준의 특성을 동시에 학습할 수 있다. MRNet은 GTZAN, FMA Small, FMA Large, Melon 데이터셋에서 각각 94.5 %, 56.6 %, 63.2 %, 71.3 %의 분류 정확도를 기록하며 기존 합성곱 신경망(Convolution Neural Network, CNN) 기반 접근법들을 능가하는 성능을 보였다. 이러한 결과는 MIR 과제에서 강건하고 계층적인 음악 표현 학습을 위한 MRNet의 효과성을 입증한다.
핵심용어: 딥러닝, 음악 정보 검색, 합성곱 신경망 (Convolution Neural Network, CNN), 음악 표현

† **Corresponding author:** Ha-Jin Yu (hjyu@uos.ac.kr)
Department of Computer Science, University of Seoul & 163, Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Republic of Korea
(Tel: 82-2-6490-2448, Fax: 82-2-6490-5697)

I. Introduction

The explosive growth of the digital music industry over the past decade has led to the release of vast amounts of new music every year.[1] This surge has driven significant interest in technologies that can automatically analyze and organize large-scale music collections.[2] One such field is Music Information Retrieval (MIR), which aims to extract high-level semantic information from audio signals, including genre, artist identity, tempo, and mood. These extracted representations serve as the foundation for various downstream tasks such as classification, recommendation, and music generation.[3,4]

Deep Neural Networks (DNNs) have shown remarkable success across many domains due to their strong capacity for data-driven feature learning. In MIR, DNNs have become increasingly dominant, with Convolutional Neural Networks (CNNs) in particular proving to be well-suited for learning from raw or spectrogram-based audio inputs.[5-8] This suitability arises from the hierarchical nature of music: short-term features like pitch, rhythm, and timbre accumulate over time to form higher-level patterns such as melody, emotional tone, or structure.[9] CNNs inherently exploit this compositionality by progressively aggregating local features into global representations through stacked convolutional layers.

Motivated by these characteristics, many prior works have designed CNN-based MIR models that capture either local or global musical traits. However, recent findings suggest that effective music representations are not confined to deep layers alone. Previous works demonstrated that shallow-layer features can also carry rich musical information, highlighting the importance of incorporating multiple levels of abstraction in MIR models.[10,11] Building on these insights, we identify two key considerations for designing an effective MIR system: (i) the ability to capture features across a wide range of time scales, and (ii) the ability to utilize representations from different processing depths.

To address both, we propose the Multi-Route Neural Network (MRNet), a CNN-based architecture specifically designed to extract music representations at various temporal resolutions and abstraction levels. MRNet is constructed by stacking Res2Blocks[12] with varying dilation rates along the time axis, allowing each block to specialize in a different temporal context. We further introduce the multi-route Res2Block (MRBlock), an enhanced module that splits the feature extraction path into three branches. Each branch is processed at a different depth, enabling parallel extraction of low-, mid-, and high-level features from a shared input.

We conduct evaluations of MRNet using four widely used datasets: GTZAN ,[13] FMA Small, FMA Large[14] and Melon Playlist,[15] with an emphasis on music classification tasks such as genre prediction. Through these experiments, we demonstrate that MRNet outperforms conventional CNN-based architectures and effectively learns hierarchical, multi-scale representations suitable for MIR.

II. Related Works

CNNs have been widely adopted in MIR due to their ability to model hierarchical patterns in time–frequency representations.[5,11] Many prior studies have explored CNN-based architectures tailored to various MIR tasks. For example, CNNs have been applied to timbre classification,[8] music tagging,[7] and genre classification across multiple datasets.[6,10,16] CNN-based models have also shown promise in music recommendation systems by learning user preference aligned representations from audio content.[9]

A key reason for CNNs' popularity in MIR is their effectiveness in capturing both local and global acoustic features through progressive convolutional layers.[5] This aligns well with the hierarchical nature of music, where short-term elements such as pitch or rhythm combine over time to form long-term patterns like melody and mood.[2,4] However, many existing CNN architectures treat deep-layer features as the primary source of semantic

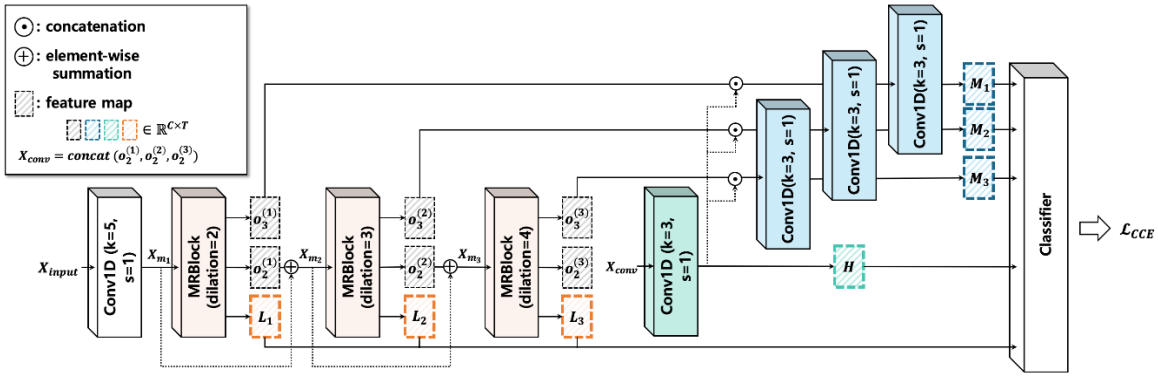

Fig. 1. (Color available online) The overall architecture of the proposed Multi-Route Neural Network. The feature map X_{conv} is constructed by concatenating the outputs $o_2^{(1)}$, $o_2^{(2)}$, $o_2^{(3)}$.

information, often overlooking the potential of shallow-layer representations.

Recent work by Liu *et al.*[10] challenged this assumption by showing that shallow features can also carry discriminative information in MIR tasks. Their findings highlight the need for architectures that integrate multi-level processing and adapt to varying temporal resolutions.

III. Proposed Method

Our goal is to design a neural architecture capable of learning music representations across both diverse temporal ranges and multiple abstraction levels. To this end, we propose the Multi-Route Neural Network (MRNet), a CNN-based framework that captures features from low, mid, and high processing depths while also modeling temporal information at multiple resolutions.

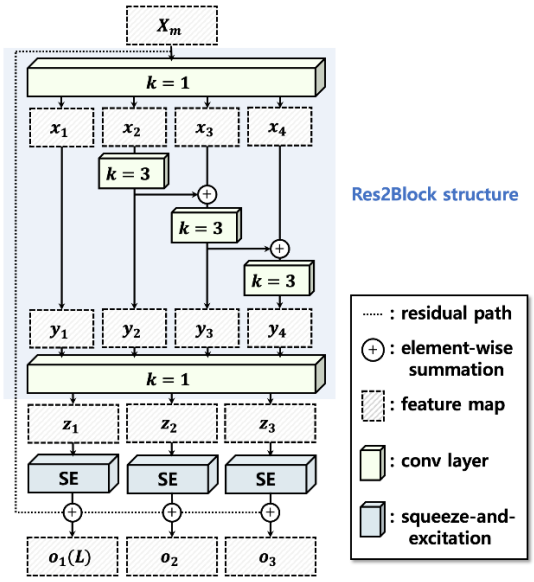

Fig. 2. (Color available online) Description of our proposed multi-route Res2Block. k means kernel size of convolution. The blue-highlighted region indicates the original Res2Block.

3.1 Overall Architecture

Fig. 1 presents the overall structure of MRNet. MRNet comprises three MRBlocks, five Convolutional layers, and a classifier. A detailed description of MRBlock, a variant of the Res2Block, is provided in Fig. 2 and the following subsection. To capture temporal structures of varying durations, the MRBlocks are stacked with increasing dilation rates (2, 3, and 4), allowing each block to specialize in a different temporal context. We explored various configurations and numbers of stacked blocks and found that this design consistently yielded the best performance across our experiments.

Given an input feature map, the first convolutional layer produces the input to the first MRBlock. For each MRBlock, the input is either the initial convolution output (when $i = 1$) or the element-wise sum of the previous MRBlock's second output $o_2^{(i-1)}$ and its own input:

$$X_{m_i} = \begin{cases} \mathrm{Conv1D}(X_{input}), & i = 1 \\ X_{m_{i-1}} + o_2^{(i-1)}, & i > 1 \end{cases}. \quad (1)$$

Each MRBlock produces three outputs $o_1^{(i)}$, $o_2^{(i)}$, $o_3^{(i)}$. These are generated from three independent SE layers without hierarchical dependency, and the indices are assigned sequentially for notational clarity. The input of MRBlock, X, is split into segment x_i, where each x_i is processed with residual convolution to produce y_i. The resulting y_i features are further transformed into z_i and passed through SE modules to generate the final outputs $o_1^{(i)}$, $o_2^{(i)}$, $o_3^{(i)}$. These outputs are used to extract representations at different depths: Low-level features (L_1, L_2, L_3 in Fig. 1) are directly taken from $o_1^{(i)}$ without further processing. High-level feature (H) is computed by concatenating $o_2^{(1)}$, $o_2^{(2)}$, $o_2^{(3)}$ and applying an additional convolution layer as:

$$X_{conv} = \mathrm{concat}\left(o_2^{(1)}, o_2^{(2)}, o_2^{(3)}\right). \tag{2}$$

$$H = \mathrm{Conv1D}(\mathrm{X_{conv}}). \tag{3}$$

Mid-level features (M_1, M_2, M_3) are generated by adding the global context H to each $o_3^{(1)}$, $o_3^{(2)}$, $o_3^{(3)}$ and passing the result through a separate convolution:

$$M_i = \mathrm{Conv1D}\left(\mathrm{o_3^{(i)} + H}\right). \tag{4}$$

This architectural design allows MRNet to extract low-, mid-, and high-level features in parallel, offering diverse representations for downstream classification.

3.2 Multi-Route Res2Block (MRBlock)

MRBlock is the core component that enables MRNet to extract multi-depth features. Inspired by findings that shallow-layer features can be effective in MIR tasks, MRBlock explicitly separates feature processing into three branches.

As illustrated in Fig. 2, the input feature map is first passed through a convolution layer and split into four segments ($y_1, \ldots, y_4$) along the channel axis. Each segment is processed independently and then recombined. The output is subsequently divided into three segments corresponding to $o_1^{(i)}$, $o_2^{(i)}$ and $o_3^{(i)}$, and each is individually refined using a Squeeze-and-Excitation (SE) mechanism. This split-aggregate-refine process allows each output to emphasize a different level of representation while sharing the same temporal scope.

Table 1. Detailed structure of the classifier. ASP denotes Attentive Statistics Pooling, which is applied for global pooling along the temporal axis. BN refers to 1D batch normalization.

Name	Layer	Input	Output
Classifier	ASP × 7	(7 × C × T)	(7 × C)
	W	(7 × C)	(7 × C)
	BN & Linear	(7C)	# of class
ASP	Conv (k = 1, s = 1)	(C × T)	(C / 16 × T)
	Conv (k = 1, s = 1)	(C / 16 × T)	(C × T)
	Softmax	(C × T)	(C × T)
	Statistics pool	(C × T)	(C × 2)

3.3 Classifier

MIR tasks vary in nature, and different target types (e.g., genre, mood) may rely more heavily on features from specific abstraction levels. To accommodate this, MRNet includes a classifier that adaptively weights and integrates the extracted features ($L_1, \ldots, L_3, M_1, \ldots, M_3, H$).

As shown in Table 1, each of the seven features is first passed through an Attentive Statistics Pooling (ASP) layer, which summarizes the temporal sequence into a fixed-length vector using attention-weighted mean and standard deviation. These vectors are then scaled by a learnable weight vector $W \in \mathbb{R}^{7\times1}$, and passed through a linear classifier. This design enables the network to learn task-dependent importance across multiple feature depths.

IV. Experiment

4.1 Dataset

We evaluated MRNet on four publicly available music datasets commonly used in music information retrieval tasks. GTZAN[13] contains 1,000 audio tracks categorized

into 10 distinct genres. Each track has a fixed duration of 30 seconds. It is widely used as a benchmark for genre classification. We also used both the small and large subsets of the Free Music Archive (FMA)[14] dataset. The small subset consists of 8,000 samples evenly distributed across 8 genre categories. The large subset contains 106,574 tracks labeled with 161 genre categories, with highly imbalanced class distributions. Melon Playlist[15] dataset includes 649,091 songs accompanied by metadata such as artist, album, and genre. For our experiments, we clustered tracks based on their genre annotations to create a genre classification task. Note that only spectrograms were available in this dataset; raw audio waveforms were not provided. For dataset splits, we followed the official training and evaluation protocols provided with the FMA datasets. For GTZAN and Melon dataset, we applied 10-fold cross validation following common practice in prior research. To facilitate reproducibility, we have also released the implementation of our K-fold splitting procedure on GitHub.[1)]

4.2 Experiment setting

We used classification accuracy as the primary evaluation metric across all datasets. For the FMA Large and Melon datasets, which contain highly imbalanced genre distributions, we additionally report the macro-averaged F1-score to better reflect performance across classes.

To obtain input features, we employed WavLM[17] as a pretrained feature extractor for the GTZAN and FMA datasets. WavLM has demonstrated strong performance in various audio-related tasks, outperforming traditional hand-crafted features. Each input waveform was transformed into frame-level embeddings using the base version of WavLM.

For the Melon dataset, since raw waveforms were unavailable, we extracted 48-dimensional Mel-spectrograms directly from the provided data. WavLM was not used for this dataset. In training process, we used the AdamW optimizer with a weight decay of 10^{-4}. The learning rate was initialized at 10^{-3} and decayed to 5×10^{-4} following a cosine annealing schedule. Models were trained for 1000 epochs with a mini-batch size of 48. All experiments were conducted on two NVIDIA A5000 GPUs. The full experimental code is available on GitHub.

4.3 Results

Comparison with Previous Works. To assess the competitiveness of the proposed MRNet, we conducted music genre classification experiments across four benchmark datasets. The results are summarized in Table 2. In addition to accuracy, we report macro F1-scores for the FMA Large and Melon datasets, which exhibit significant class imbalance. MRNet achieved top performance on all four datasets, recording classification accuracies of 94.5 %, 56.6 %, 63.2 %, and 71.3 % on GTZAN, FMA Small, FMA Large, and Melon, respectively. Notably, the gains on FMA Large and Melon are particularly

Table 2. Experimental results across various datasets. Macro F1-scores are reported only for unbalanced datasets to complement the accuracy metric.

Models	Dataset	Samples	Accuracy (%)	Macro-f1
MoER[18]	GTZAN	1,000	86.4	N/A
BBNN[10]			93.9	N/A
Siddiquee *et al.*[19]			90.0	N/A
MRNet (ours)			**94.5**	N/A
MoER[18]	FMA (Small)	8,000	55.9	N/A
BBNN[10]*			54.8	N/A
LFCNet[16]			55.1	N/A
MRNet (ours)			**56.6**	N/A
BBNN[10]*	FMA (Large)	106,574	53.9	0.34
LFCNet[16]*			52.7	0.35
MRNet			**63.2**	**0.38**
ResNet34	Melon	649,091	63.6	0.36
SE-ResNet34			64.1	0.38
BBNN[10]*			60.2	0.36
LFCNet[16]*			64.7	0.41
MRNet (ours)			**71.3**	**0.55**

* indicates results from our implementation.

1) https://github.com/Jungwoo4021/MRNet

Table 3. Experiment results of route ablation experiments on the FMA-Small dataset.

L_1	L_2	L_3	M_1	M_2	M_3	H	Acc (%)
							56.6
×							53.9
	×						51.9
		×					54.5
			×				54.8
				×			54.6
					×		53.0
						×	53.0

significant, suggesting that MRNet is well-suited for handling datasets with large-scale, fine-grained class structures. These results demonstrate MRNet's superior generalization capability and robust clustering of musical characteristics, even in complex and diverse music collections.

Ablation Study. To validate the contribution of the multi-route architecture, we conducted a route ablation study using the FMA Small dataset. Table 3 shows the performance degradation when individual routes were disabled.

The second row in the table presents the baseline performance of the full MRNet model, which utilizes all seven feature paths. Rows 3 to 9 depict the results when one of the feature branches (e.g., L_1, M_2, H, etc.) was removed during training and evaluation. In all configurations, performance dropped below the original 56.6 % accuracy of the full model. These findings confirm that each route contributes meaningfully to the model's discriminative power and that the full multi-path design is essential for optimal performance.

Feature utilization by task. MRNet extracts seven feature representations: three from low-level (L, ... , L_3), three from mid-level (M_1, ... , M_3), and one from high-level (H) branches. Depending on the task, the importance of each feature type may vary. For example, mood classification may rely more on long-term global features, while genre classification might benefit from

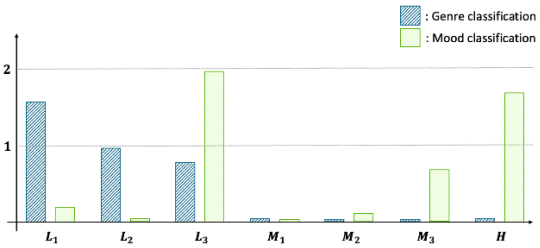

Fig. 3. (Color available online) Task-specific distribution of the learned feature scaling vector W for genre and mood classification.

short-term local patterns.

To explore this, we examined the learned weights of the feature scaling vector W, introduced in Section 3.3, Fig. 3 illustrates the distribution of W values after training MRNet for both genre and mood classification tasks.

The analysis shows that low-level features, especially L_1 and L_2, were most influential in genre classification, supporting prior findings that shallow features are highly effective for this task. In contrast, mood classification placed more weight on L_3, M_3, and H features, which have deeper and broader temporal receptive fields. These results confirm that MRNet dynamically adjusts its feature emphasis according to the target MIR objective, enhancing task-specific performance.

Although the contribution of M_1, M_2, M_3 and H appeared relatively weak in the FMA Small dataset, Table 3 shows that removing mid-level features consistently degraded performance, indicating that these branches are not redundant and do contribute to the overall effectiveness of MRNet.

V. Conclusions

In this paper, we introduced MRNet, a novel convolutional architecture tailored for MIR. MRNet is designed to capture musical representations across multiple time scales and processing depths by employing distinct feature extraction routes. Leveraging a stack of dilated Res2Blocks and the proposed multi-route Res2Block (MRBlock), MRNet effectively extracts low-,

mid-, and high-level features in parallel.

Through extensive evaluations on four benchmark datasets including GTZAN, FMA Small, FMA Large, and Melon Playlist, we demonstrated that MRNet consistently outperforms previous CNN-based approaches in genre classification tasks. In addition, our analysis of the learned feature scaling weights revealed that MRNet can dynamically prioritize different types of features depending on the specific MIR objective, such as genre or mood classification.

These results highlight MRNet's ability to learn robust, hierarchical representations that are both flexible and task adaptive. As part of future work, we plan to extend MRNet to other MIR tasks beyond classification, including playlist recommendation, emotion prediction, and music generation, to further explore its generalization and compositional capabilities. Furthermore, we intend to analyze the differences between misclassified and correctly classified tracks to better understand the limitations of MRNet.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2023R1A2C1005744).

References

1. G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans on speech and audio process. **10**, 293-302 (2002).
2. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," Proc. IEEE, **96**, 668-696 (2008).
3. D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, P. H. Boyer, O. Mayor, G. R. Trepat, J. Salamon, J. R. Z. González, and X. Serra, "Essentia: An audio analysis library for music information retrieval," Proc. 14th ISMIR, 493-498 (2013).
4. A. Van Den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," Proc. NIPS, 26 (2013).
5. K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," Proc. ICASSP, 2392-2396 (2017).
6. Y. Xu and W. Zhou, "A deep music genres classification model based on CNN with Squeeze & Excitation Block," Proc. IEEE APSIPA ASC. 332-338 (2020).
7. A. Ferraro, D. Bogdanov, X. S. Jay, H. Jeon, and J. Yoon, "How low can you go? Reducing frequency and time resolution in current CNN architectures for music auto-tagging," Proc. 28th EUSIPCO, 131-135 (2021).
8. D. Kim, T. T. Sung, Y. S. Cho, G. Lee, and B. C. Sohn, "A single predominant instrument recognition of polyphonic music using CNN-based timbre analysis," Int. J. Eng. Technol. **7**, 590-595 (2018).
9. S. Joshi, T. Jain, and N. Nair, "Emotion based music recommendation system using LSTM-CNN architecture," Proc. IEEE ICCCNT. 01-06 (2021).
10. C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu, "Bottom-up broadcast neural network for music genre classification," Multimed. Tools Appl. **80**, 7313-7331 (2021).
11. J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," IEEE Signal Processing Letters, **24**, 1208-1212 (2017).
12. S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," IEEE Trans. Pattern Anal. Mach. Intell. **43**, 652-662 (2019).
13. I. Ikhsan, L. Novamizanti, and I. N. A. Ramatryana, "Automatic musical genre classification of audio using Hidden Markov Model," Proc. 2nd ICoICT, 397-402 (2014).
14. *FMA: A Dataset for Music Analysis*, https://arxiv.org/abs/1612.01840, (Last viewed September 16, 2025).
15. A. Ferraro, Y. Kim, S. Lee, B. Kim, N. Jo, S. Lim, S. Lim, J. Jang, S. Kim, and X. Serra, "Melon playlist dataset: A public dataset for audio-based playlist generation and music tagging," Proc. IEEE ICASSP, 536-540 (2021).
16. S.-H. Cho, Y. Park, and J. Lee, "Effective music genre classification using late fusion convolutional neural network with multiple spectral features," Proc. IEEE ICCE-Asia, 1-4 (2022).
17. S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian , Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," IEEE J.

Sel. Top. Signal Process. **16**, 1505-1518 (2022).
18. Y. Yi, K.-Y. Chen, and H.-Y.Gu, “Mixture of CNN experts from multiple acoustic feature domain for music genre classification,” Proc. IEEE APSIPA ASC. 1250-1255 (2019).
19. Md. N. A. Siddiquee, Md. A. Hossain, and F. Wahida, “An effective machine learning approach for music genre classification with mel spectrograms and Knn,” Proc. IEEE IC3S, 1-4 (2013).

Profile

▸ Jungwoo Heo (허 정 우)

He received his B.S. degree in Computer Science in 2022 from the University of Seoul, Seoul, South Korea, where he is currently pursuing a Ph.D. degree.
His research interests include speaker recognition, audio spoofing detection, music information retrieval, and deep learning.

▸ Hyun-seo Shin (신 현 서)

He received his B.S. degree in Computer Science in 2018 from the University of Seoul, Seoul, South Korea, where he is currently pursuing a Ph.D. degree.
His research interests include speaker recognition, audio deepfake detection, music information retrieval, and deep learning.

▸ Chan-yeong Lim (임 찬 영)

He received his B.S. degree in Statistics in 2023 from the University of Seoul, Seoul, South Korea, where he is currently pursuing an M.S. degree.
His research interests include speaker recognition, audio spoofing detection, model compression, and deep learning.

▸ Kyo-won Koo (구 교 원)

He received his B.S. degree in the School of Electrical and Computer Engineering in 2024 from the University of Seoul, Seoul, South Korea, where he is currently pursuing an M.S. degree.
His research interests include speaker recognition, speaker diarization, and deep learning.

▸ Seung-bin Kim (김 승 빈)

He received the B.S. degree in computer science in 2020 from the University of Seoul, Seoul, South Korea, where he is currently working toward the M.S. degree.
His research interests include speaker recognition, audio spoofing detection, and deep learning.

▸ Jisoo Son (손 지 수)

He received his B.S. degree in Computer Science in 2025 from the University of Seoul, Seoul, South Korea, where he is currently pursuing an integrated M.S./Ph.D. degree.
His research interests include speaker recognition, audio classification, audio deepfake detection, and defense against adversarial attacks.

▸ Ha-Jin Yu (유 하 진)

He received his B.S., M.S., and Ph.D. degrees in Computer Science from KAIST, South Korea, in 1990, 1992, and 1997, respectively.
From 1997 to 2000, he was a Senior Researcher at LG Electronics, and from 2000 to 2002, he was the Director at SL2 Ltd. Since 2002, he has been working as a Professor in the Depart-ment of Computer Science, University of Seoul.
His research interests include speech and speaker recognition, audio spoofing detection, music information retrieval, and machine learning.
He is an Editor of the Acoustical Society of Korea.