NEURAL AUDIO FINGERPRINTING

Jongsoo, Kim
GSCT, KAIST
jongsoo.kim@kaist.ac.kr

Hyunjae, Kim
GSCT, KAIST
present@kaist.ac.kr

Hail, Song
GSCT, KAIST
hail96@kaist.ac.kr

ABSTRACT

Although audio fingerprinting systems has been developing, it still has a limitation to find correct songs when distorted audios is input. Audio inputs can be distorted in various forms depending on the usage of system, requiring a robust audio fingerprinting system that can be applied in all applications. In addition, the performance of the audio fingerprinting system is evaluated complexly by various criteria: robustness, reliability, granularity, fingerprint size, and search speed. So, it is a really difficult task to implement an audio fingerprinting system that satisfies all of these criteria. In order to overcome these limitations, the audio fingerprinting systems using deep-learning are suggested recently such as Separable CNN, Attention-base model. Among them, the contrastive learning method is optimized for audio fingerprinting system. And some researchers tried to summarize information about the time and frequency axes of the audio separately for extracting the most distinctive features. In this project, audio fingerprinting systems using deep learning were directly implemented and analyzed the performance of several models in terms of various criteria.

1. INTRODUCTION

Audio fingerprinting is one of the main technique in the music information retrieval field. Like human fingerprinting that everyone has unique to each other, audio fingerprinting is a unique mark corresponding to a given audio data. Using this technology, the audio in the type of waveform can be handled in a compressed form, making it more cost effective when comparing huge amounts of audio data, such as music search. In fact, audio fingerprinting is essential technique for music search services such as Shazam [1]. Conventionally, audio fingerprints have been composed by extracting energy peak values of the Mel-spectrogram [3, 9]. However, these methodologies have a problem in that the performance is drastically deteriorated in the noisy environment or with severely distorted audio. As a breakthrough, neural audio fingerprinting technique with contrastive learning methods is being utilized in recent research [4]. Audio Fingerprinting can be trained and generated directly through a deep learning model only with audio source data; without time-consuming labeling. Generally, deep learning models for neural audio fingerprinting are consist of encoder block and projection block with Mel-spectrogram data of audio as input [5, 8]. In some previous studies, the neural network consisting of separable convolution layers was used to alternately process each axis of 2-D data [4, 5]. Because, the both axis of Mel-spectrogram, time and frequency axis, are contains important information to recognize the unique characteristics of audio source. On the other hand, the attention mechanism for time and frequency axis is adopted before input of encoder [8, 10]. All of these approaches tried to effectively consider the time and frequency information of the Mel-spectrogram. In this study, we tried to implement various models including previous models, and see how these audio fingerprinting works for various environments such as the formal dataset or actual streaming sound source. Especially, we compared and discussed the results of these several experimental trial in terms of important criteria for audio fingerprinting performance: Robustness, Reliability, Granularity, and Fingerprint size.

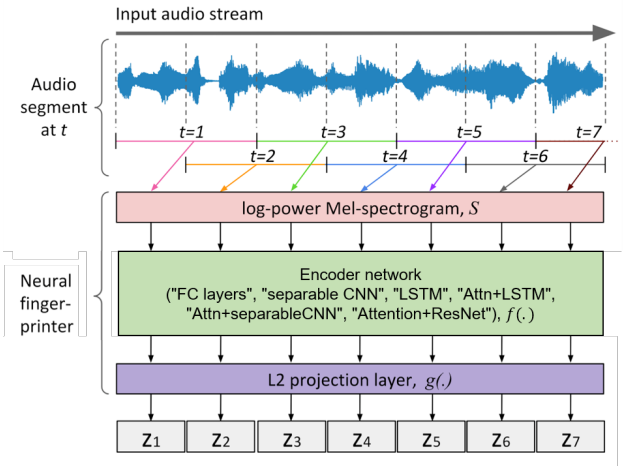

Figure 1. Overview of the neural audio fingerprint system. Encoder network f(.) and L2 projection layer perform audio fingerprinting task.

2. METHODS

The overall structure of the system is shown in Figure 1. First, the input audio stream is divided into chunks by a unit of time and becomes an audio segment. It is then converted to log-power Mel-spectrogram S. Finally, the data

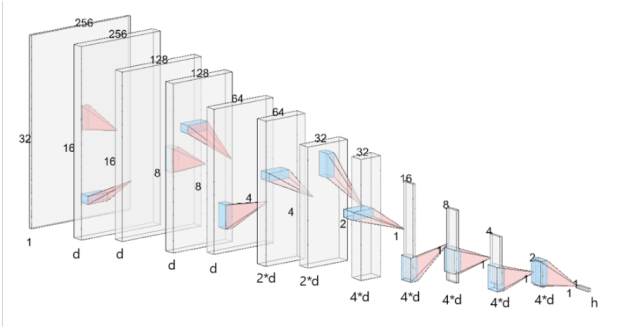

Figure 2. Illustration of separable CNN structure.

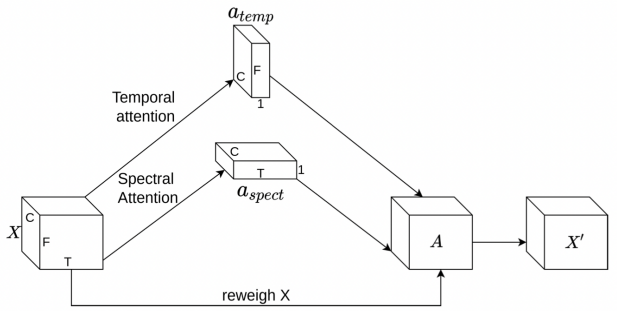

Figure 3. Illustration of attention structure. [8]

are converted into fingerprints of d-dimensional vectors via Encode network $f(.)$ and L2 projection layer $g(.)$. We compared the performance by diversifying the encoder network structure in this neural audio fingerprinting process. In addition, to verify the performance of the L2 projection, some network structures have replaced the $g(.)$ portion with a fully connected layer. Encoder networks used the following six types.

- Fully connected layers
- Separable CNN
- LSTM layer
- Attention layer + LSTM layer
- Attention layer + separable CNN
- Attention layer + ResNet

We benchmark [4] as a baseline of the overall network structure. The description of each method is as follows.

2.1 Contrastive Learning

All our models used constrative learning. Contrastive prediction is a methodology that enables robust classification work on replica data. We used this method for the robustness of audio fingerprinting. The loss function for learning the pairwise similarity of the replica and the original data is as follows.

$$l(i,j) = -log(\frac{exp(a_{i,j}/\tau)}{\sum_{k=1}^{N} \mathbb{1}(k \neq i) exp(a_{i,j}/\tau)}) \quad (1)$$

$$L = \frac{1}{N} \sum_{k=1}^{N} [l(2k-1, 2k), l(2k, 2k-1)] \quad (2)$$

2.2 Separable CNN

As in Baseline [4], the Separable CNN structure is implemented as an encoder. Separable CNN is a network that alternately configures unbalanced convolution kernels of the time and frequency axes. It has the characteristic that time and frequency features can be extracted well, especially on audio data. [7] was referred to in the configuration of the separable convolution network. The overall network structure of the separable convolution network is in Figure 2.

2.3 Fully Connected Layers

A fully connected neural network with a simple structure was also composed of an encoder to experiment. The encoder was constructed using three layers according to the size of the L2 projection. The number of nodes in the layer was (8192, 512, and 256). Additionally, the experiment was performed by replacing $g(.)$ with fully connected layers to confirm the performance of the L2 projection.

2.4 LSTM

A mel-spectrogram composed of time and frequency axes has the characteristics of time series data. That is why we considered long short term memory layer, which learns by remembering long term memory and short term memory, as a suitable candidate for encoder. The experiment was conducted using the LSTM layer as an encoder.

2.5 Attention

As shown in Figure 3, the Attention method generates two attentions based on the temporary axis and the spectral axis, and then multiplies these two attentions by a matrix. Then, in order to make the gradient descent easier, scaling is performed as Eqn (5). We conducted experiments by adding this attention structure to front of the LSTM and sCNN model. In addition, the model presented in the paper referring to this was additionally experimented.

$$A_{temp} = \text{Softmax}(X^{\top} W_{temp}) \quad (3)$$

$$A_{spect} = \text{Softmax}(X^{\top} W_{spect}) \quad (4)$$

$$X' = A_{temp} \otimes A_{spect} \times S \quad (5)$$

3. EXPERIMENTAL SETUP

3.1 Dataset

We used the following two dataset for the main experiments. Due to limitations in downloading many songs from YouTube, we evaluated about only 100 songs, and for fair comparison, FMA dataset also used only 100 in the evaluation.

- **Free Music Archive (FMA) Dataset** We used the FMA medium dataset for training(10k), validation(0.5k), and evaluation(0.1k) (30s).

- **YouTube (0.1k)** We used the real music dataset which downloaded from YouTube. This dataset was used in only evalation.

3.2 Augmentation

Unlike the referenced papers, for easy control, we used the 'torchaudio-augmentations' package instead of augmentaion dataset. Table 1 is the configuration for data augmentation.

Type	Parameter	Value
Noise	SNR {min, max}	{0.1, 0.3}
Delay	Delay {min, max}	{100, 300} ms
Reverb	Reverberance {min, max}	{90, 91}
	Room Size {min, max}	{90, 91}

Table 1. Configuration for data augmentation

3.3 Implementation Details

We used the Mel-Spectrogram as an input in our experiments. Table 2 is the configuration for the Mel-Spectrogram.

Parameter	Value
Sample Rate	8,000 Hz
STFT window function	*Hann*
STFT window {length, hop}	1024, 256
n_mels	256
Frequency {min, max}	{300, 4000} Hz
Fingerprint {window length, hop}	{1, 0.5} *s*
Fingerprint dimension, *d*	64 or 128
Batch Size, *N*	640

Table 2. Shared configurations for experiments

3.4 Criteria

For audio search task, not only the audio fingerprint extraction itself, but also the accuracy, efficiency, and computational structure of algorithm are important indices for evaluation of audio fingerprinting [3, 6]. However, since this project does not mainly investigated about search phase, we indirectly assessed about the following criteria in our experiments.

- **Robustness:** Can we find the exact audio clips in various noise environment?
- **Reliability:** How often does the model make false inferences?
- **Granularity:** How short is the query length to find audio clips?
- **Fingerprint Size:** How much memory storage does the fingerprint occupy?

		Value		
Type	Parameter	Low	Mid	High
Noise	SNR min	0.0	0.0	0.0
	SNR min	0.2	0.3	0.4
Delay	Delay min	50	80	200
	Delay max	60	90	300
Reverb	Reverberance min	61	91	100
	Reverberance min	61	91	100
	Room Size max	50	90	99
	Room Size max	60	91	100

Table 3. Configurations for evaluation

3.5 Evaluation

We used the Top-1 Hit Rate for evaluation as follow:

$$accuracy = \frac{n_{hits}@top-1}{(n_{hits}@top-1)+(n_{miss}@top-1)} \quad (6)$$

And there are 3 matching cases.

- **Exact:** How much our system finds the correct index of songs.
- **Near:** Match within $\pm$ 1 index or $\pm$ 500 ms.
- **Correct:** How much our system finds songs correctly.

4. RESULTS AND DISCUSSION

4.1 Validation Accuracy

In general, the model which had consider the both time axis and frequency axis shown higher evaluation performance, such as Attn+Residual+L2, LSTM+L2, or sCNN+L2. However, if we attached attention layer in front of sCNN or LSTM, the output results are degraded. On the other hand, fully connected model showed incomparably worse accuracy. About this result, we suspect that the flattened input interfere the proper processing the time and frequency information of Mel-spectrogram. Therefore, the results of fully connected layer are excluded in the following evaluation results.

4.2 Fingerprint Size & Distortion Level

To evaluate the effect of fingerprint size, we conducted two different fingerprint dimension; 64 and 128. In addition, to assess the model robustness, we tested with three different distortion level for evaluation dataset. As shown in the Table 4, the proper dimension size of fingerprint is largely depend on the model structure and regarding noise level. However, as a whole, Attn+ResNet model yields outstanding performance, probably because attention layer well dealt with time and frequency information of Mel-spectrogram and residual connection can consistently deliver this information along with the whole model stream.

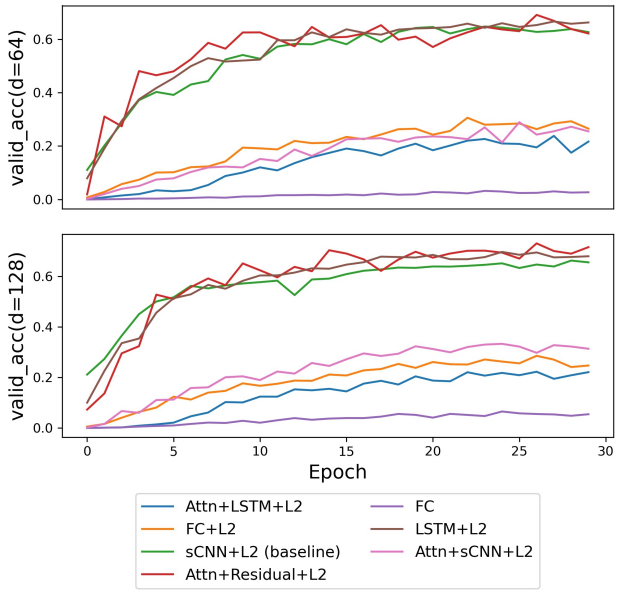

Figure 4. Validation accuracy over epoch according to fingerprint dimension

d	64			128		
Level	Low	Mid	High	Low	Mid	High
sCNN	0.8	0.85	0.8	0.9	0.8	0.65
LSTM	0.9	0.75	0.75	0.85	0.75	0.8
Attn+ sCNN	0.75	0.7	0.75	0.65	0.8	0.85
Attn+ LSTM	0.6	0.75	0.75	0.55	0.75	0.85
Attn+ ResNet	0.85	0.95	0.8	0.95	0.9	0.9

Table 4. Top-1 hit rate of FMA_medium (0.1k songs) segment-level search. d is the dimension of fingerprint embedding. 'Level' is the distortion level. (Query Length = 1s, *correct* matching)

4.3 Query Length

To check the granularity, we conducted test experiment with 7 different length of query; 1, 2, 3, 5, 6, and 10 seconds. In Table 5, as we can easily expect, the longer query length give better performance to identify the music clips for all model structures. For a short length range, 1, 2, and 3 seconds, Attn + ResNet model outperformed. However, in the longer length, pure CNN or LSTM based model shown better performance, which may indicates that the attention mechanism can help to process short data effectively, but when enough data comes in, it tends to impair some adequate information.

4.4 Real Dataset

Lastly, we tested different aspect of robustness, which is consistency for different audio system environments not noise environments. We used totally different type of music clips, collected in Youtube, from FMA dataset. The fingerprint size is 64 with 1 second query length, and noise level is 'High'. As shown in Table 6, of course, the FMA test data were corretly identified than Youtube data in general. However, despite the fact that the data type has never been trained, Attn+LSTM model can reach out to 55% accuracy, which means that, to some extent, it can operate robustly on totally different type of music clip data.

Model	Query Length (s)					
	1	2	3	5	6	10
sCNN	0.6	0.65	0.75	0.85	0.95	0.95
LSTM	0.5	0.75	0.75	0.8	0.85	0.95
A+sCNN	0.55	0.7	0.8	0.85	0.9	0.9
A+LSTM	0.4	0.65	0.7	0.85	0.85	0.9
A+ResNet	0.65	0.75	0.8	0.85	0.9	0.9

Table 5. Top-1 hit rate performance in the segment-level search for varying query lengths. In model types, 'A' denotes the Attention. (Level='Mid', d=64, *exact* matching)

	sCNN	LSTM	Attn+ sCNN	Attn+ LSTM	Attn+ ResNet
YT	0.2	0.2	0.3	0.55	0.45
FMA	0.45	0.55	0.55	0.65	0.65

Table 6. Top-1 hit rate performance for varying dataset. 'YT' denotes the dataset consisting of songs that we downloaded from YouTube. (d=64, Level='High', Query Length=1s, *near* matching)

5. CONCLUSIONS AND FUTURE WORK

In summary, we investigate the performance of audio fingerprint in the aspect of important criteria with various conditions. From our results, we validate that properly handling the time and frequency information of audio data through specific model structure, such as seperable convolution layers or attention layers, is crucial factor of generating the robust fingerprinting. And also, we experimentally assess the effect of query length and dimensionality of fingerprint for granularity and fingerprint size criteria. For future study, we can additionally deal with the later procedure of encoder of model, like projection block, building database, clip matching process, and so on. In the case of projection, we can think about the binary projection into Hamming space as studied in [2, 11]. And for database building, large DB class make model difficult to identify the correct music clips, so the effective DB composition and searching methods should be regarded. Moreover, it may be possible to try and study about audio fingerprinting with precedes melody extraction to cover the contents such as cover music and remixes [2, 11].

6. CONTRIBUTION OF EACH TEAM MEMBER

- **Jongsoo, Kim [20223166]:** Configuring about the models which used the Attention and the Fully Connected Layers. Data pre-processing. Training the models. Evaluation. Writing report (all the parts).
- **Hyunjae, Kim [20225090]:** Configuring the baseline model, Model training, Writing and revision report (Result and Discussion, Criteria)
- **Hail, Song [20223342]:** Configuring the LSTM Model Structure, Visualization of learning results, Drawing model structure figures, Model training, Writing report (Method, Setup)

7. REFERENCES

[1] Shazam music recognition service.

[2] Abraham Báez-Suárez, Nolan Shah, Juan Arturo Nolazco-Flores, Shou-Hsuan S. Huang, Omprakash Gnawali, and Weidong Shi. Samaf: Sequence-to-sequence autoencoder model for audio fingerprinting. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(2), may 2020.

[3] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of audio fingerprinting. *Journal of VLSI signal processing systems for signal, image and video technology*, 41(3):271–284, November 2005.

[4] Sungkyun Chang, Donmoon Lee, Jeongsoo Park, Hyungui Lim, Kyogu Lee, Karam Ko, and Yoonchang Han. Neural audio fingerprint for high-specific audio retrieval based on contrastive learning. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, 2021.

[5] Beat Gfeller, Blaise Aguera-Arcas, Dominik Roblek, James David Lyon, Julian James Odell, Kevin Kilgour, Marvin Ritter, Matt Sharifi, Mihajlo Velimirović, Ruiqi Guo, and Sanjiv Kumar. Now playing: Continuous low-power music recognition. In *NIPS 2017 Workshop: Machine Learning on the Phone*, 2017.

[6] Jaap Haitsma and Ton Kalker. A highly robust audio fingerprinting system. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2002.

[7] Franck Mamalet and Christophe Garcia. Simplifying convnets for fast learning. In *International Conference on Artificial Neural Networks*, pages 58–65. Springer, 2012.

[8] Anup Singh, Kris Demuynck, and Vipul Arora. Attention-based audio embeddings for query-by-example. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[9] Avery Wang. An industrial strength audio search algorithm. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2003.

[10] Helin Wang, Yuexian Zou, Dading Chong, and Wenwu Wang. Environmental sound classification with parallel temporal-spectral attention. In *Interspeech*, 2019.

[11] Xinyu Wu and Hongxia Wang. Asymmetric contrastive learning for audio fingerprinting. *IEEE Signal Processing Letters*, 29:1873–1877, 2022.