Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 20 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i20.44221

PAPER

Classification and Retrieval of Multimedia Audio Learning Resources

Wenwen Zhang(✉)

School of Marxism, Hebei Academy of Fine Arts, Shijiazhuang, China

zhangwenwen@hbafa.edu.cn

ABSTRACT

With the development of the Internet and new media, multimedia and audio learning resources have been widely used in teaching and learning. However, their classification and retrieval have become important and urgent issues to be addressed. This study conducted in-depth research on the classification system, construction, and retrieval of multimedia audio learning resources, with the aim of solving several problems with existing research methods, such as time-consuming manual labeling, inconsistent labeling, and traditional retrieval methods neglecting the correlation between audio and metadata. First, a classification model of audio learning resources was constructed. It processed single-mode data from audios and annotated texts and further abstracted the single-mode information into high-level feature vectors. Then the complementarity between multi-modalities was used to fuse the abstract features or decision-making results and eliminate information redundancy between modalities, thereby learning a better feature representation of multimedia audio learning resources. Second, a retrieval method for the resources based on self-similarity matrix filtering was proposed, which aimed to improve the accuracy and efficiency of retrieval. This study provides a new theoretical and practical perspective for classifying and retrieving multimedia audio learning resources.

KEYWORDS

multimedia audio learning resources, construction of a classification model, resource retrieval, self-similarity matrix filtering, feature representation

1 INTRODUCTION

With the increasing development of global informatization in today's society, the rapid development of the Internet and the widespread use of new media have led to many multimedia audio learning resources, greatly enriching people's learning methods and means [1–17]. These resources cover various fields, from basic education to higher education and from subject teaching to vocational skill training, and exist in rich and diverse forms, such as audio explanations, teaching videos, online courses, and so on [18] [19]. With intuitive visual and auditory information as the

Zhang, W. (2023). Classification and Retrieval of Multimedia Audio Learning Resources. *International Journal of Emerging Technologies in Learning (iJET)*, 18(20), pp. 99–113. https://doi.org/10.3991/ijet.v18i20.44221

Article submitted 2023-06-17. Revision uploaded 2023-08-10. Final acceptance 2023-08-13.

medium, these resources convey knowledge and information more vividly and concretely, bringing great convenience to learners [20] [21]. However, with the rapid increase in audio learning resources, some problems have occurred in their management and utilization. It is urgent to effectively classify and retrieve these resources in order to help users obtain the required learning resources efficiently and quickly.

Furthermore, effective classification and retrieval of multimedia audio learning resources not only directly affect the efficiency of using the resources and the user experience but also play an important role in promoting the educational informatization process [22]. First, a good classification and retrieval system enables users to quickly find suitable resources among numerous resources and avoid wasting time on ineffective searches, thereby improving learning efficiency. Second, the value of resources can be fully utilized by managing the resources effectively, and the fair distribution and use of educational resources can be promoted, thereby helping promote educational equity [23] [24]. In addition, the classification and retrieval technology of multimedia audio learning resources also has a profound impact on related research fields, including audio recognition, automatic classification, information retrieval, etc. With the development of deep learning, big data, artificial intelligence, and other technologies, these resources' classification system construction and retrieval have become a research direction worthy of attention and in-depth discussion [25] [26].

Although the classification and retrieval of multimedia audio learning resources have significant practical value and research significance, existing research methods and technologies still have some shortcomings. Many research methods mainly rely on manual labeling and classification of audio resources, leading to obvious problems. On the one hand, the investment in human resources is huge, time-consuming, and inefficient, making it difficult to meet the processing needs of massive audio resources. On the other hand, the manual labeling results may lead to inconsistent labeling due to several factors, such as individual cognitive differences and knowledge backgrounds, affecting the classification accuracy [27] [28]. In addition, traditional audio retrieval methods often focus on a single audio feature only, ignoring the correlation between metadata such as audios and annotated texts. Therefore, it is often difficult for their retrieval efficiency and accuracy to achieve the ideal state [29] [30].

To address the above issues, this study conducted in-depth research on the classification system construction and retrieval of multimedia audio learning resources. The main content is divided into two parts. First, a classification model of audio learning resources was constructed that processed single-mode data from audios and annotated texts and further abstracted the single-mode information into high-level feature vectors. Then the complementarity between multi-modalities was used to fuse the abstract features or decision-making results and eliminate information redundancy between modalities, thereby learning a better feature representation of multimedia audio learning resources. Second, a retrieval method for the resources based on self-similarity matrix filtering was proposed, which aimed to improve the accuracy and efficiency of retrieval. It is expected that this study can provide a new theoretical and practical perspective for the classification and retrieval of multimedia audio learning resources and further promote the effective utilization and research of multimedia educational resources.

2 CONSTRUCTING A CLASSIFICATION MODEL OF MULTIMEDIA AUDIO LEARNING RESOURCES

The audio and annotated texts of multimedia audio learning resources already have rich feature information in their respective modalities. The information

contained in audios is mainly reflected in pronunciation, sound effects, and other aspects that reflect the language content, explanation methods, tone strength, speed, and other features of learning resources. The information in annotated texts is mainly reflected in words, symbols, and other aspects, which reflect the themes, content framework, key concepts, and other features of learning resources. These two parts of information have rich feature information in their respective modalities and are the important information carriers of audio learning resources. By processing single-mode data, valuable features can be extracted from these different information carriers, providing a foundation for subsequent classification work.

Another advantage of processing single-mode data is that information redundancy between modalities can be minimized and classification accuracy can be improved. In actual audio learning resources, a certain degree of overlap and redundancy may exist between audio and annotated texts. The redundant information may cause confusion in classification, which reduces classification accuracy. Therefore, the information redundancy between modalities can be effectively reduced by processing single-mode data and abstracting it into high-level feature vectors, there by improving classification accuracy.

Although both audio and annotated texts are information carriers for audio learning resources, the information they provide has different focuses and is complementary in some aspects. Audio information focuses more on conveying language and sound information, while annotated texts focus more on conveying structured knowledge and information. After processing the two parts of data, the complementarity between multi-modalities can be used to fuse the corresponding single-mode information, which obtains more comprehensive and representative feature representation, thereby improving the classification effect of audio learning resources.

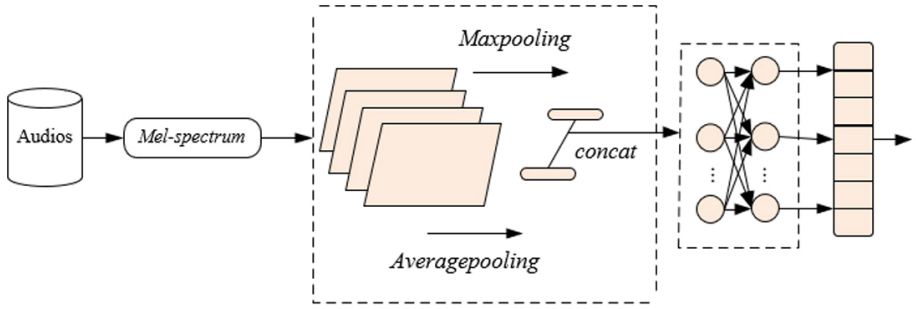

Fig. 1. Schematic diagram of audio processing process using the model

In this study, a convolutional neural network (CNN) model with maximum and average pooling was used to process the audio of multimedia audio learning resources. The core concept of this processing method is to convert audio information into image information, extract features from the image information using the CNN model, and finally obtain the audio feature vector through feature fusion.

Figure 1 provides a schematic diagram of the audio processing process using the model. The audio processing channel first processed the original audio data into a logarithmic Mel spectrogram, which aimed to convert audio information into image information. The features were extracted using the four convolutional layers with the same-size filters. Using the design of convolutional layers, not only the basic features of audio signals were extracted, but also higher-level and more abstract features were captured through stacked layers. Let $s \in R^{l \times b}$ be the two-dimensional Mel spectrogram sample, i.e., the input of the audio channel, and $j_u \in E^{3 \times 3}$ be the convolution kernel. The main architecture of this channel includes four convolutional

layers, which have 128 convolution kernels with a size of 3×3. Let s be the audio spectrogram of multimedia audio learning resource samples, l and b be the length and width of the spectrogram, A^e be the output of the last convolutional layer, and *Conv*2 be a convolutional layer. The expression of the spectrum processing process was as follows:

$$A^e = \begin{cases} \text{conv2}(s, j_u), e = 1 \\ \text{conv2}(A^{e-1}, j_u), 1 \le e \le 4 \end{cases} \tag{1}$$

The model provided a dual pooling structure to compress the effective features and reduce the calculation amount. It greatly reduced the dimensionality of features while retaining key information, thereby effectively reducing the complexity and calculation amount of the model. Let O be the output splicing results of max and average pooling, *MAX* be the global maximum pooling layer, *AV* be the global average pooling layer, and *CON* be the feature splicing layer. Then there were:

$$O = CON(MAX(A^e), AV(A^e)), e = 4 \tag{2}$$

The final audio feature vector was obtained using two dense layers. This design achieved a nonlinear combination of features and enhanced the model's expression ability. Using these two dense layers, the model learned deeper and more abstract features in audio data and obtained more effective feature representations. Let S be the output features of the sample audio channel, Q_1^y be the weight transposition of the dense layer, and n_1 be the offset. If the ReLU function was used as the activation function of the dense layer, then there were:

$$S = \text{ReLU}(Q_1^y O + n_1) \tag{3}$$

$$\text{ReLU}(x) = \begin{cases} z, if & z > 0 \\ 0, if & z \le 0 \end{cases} \tag{4}$$

In this study, the Bidirectional Encoder Representations from Transformers (BERT) model was used to process the annotated texts of multimedia audio learning resources. The basic idea of this processing method is to embed annotated texts into the token sequence of the model, extract the feature vector of a specific token as the feature representation of annotated texts, and finally obtain the music style class through feature fusion.

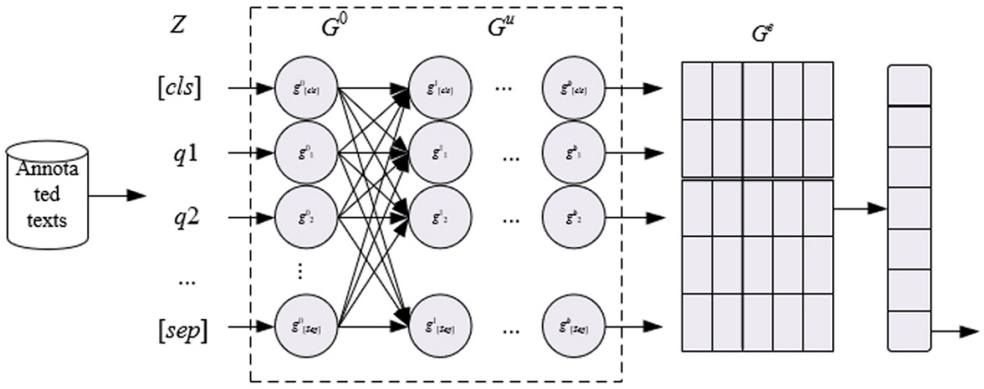

Fig. 2. Schematic diagram of annotated text processing process using the model

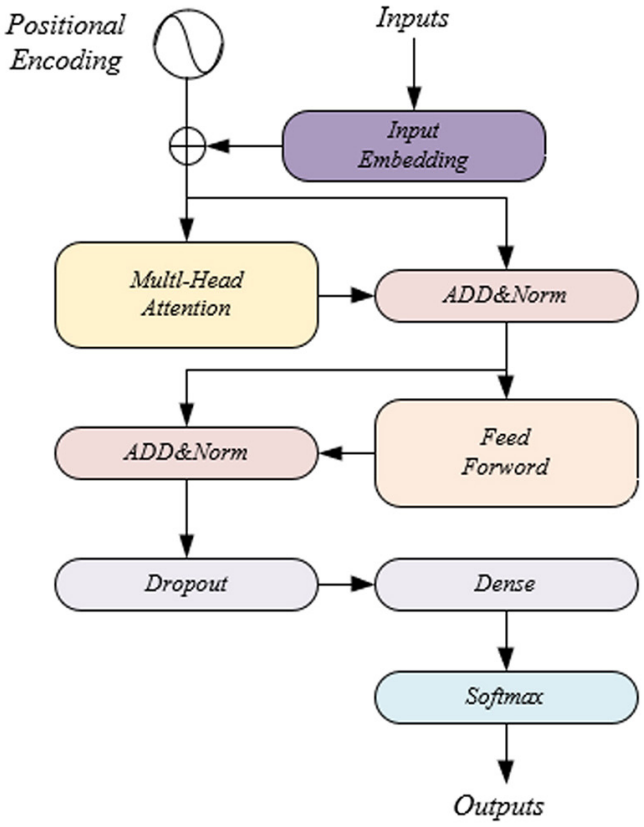

Fig. 3. Diagram of text channel structure

Figure 2 shows the schematic diagram of the annotated text processing process using the model. The annotated text processing channel first embedded annotated texts into the token sequence of the BERT model, which, besides converting texts into numerical representations for computer processing, also captured complex structural and semantic relationships in the texts. Then the model extracted the feature vector of a specific token from the token sequence as the feature representation of annotated texts. Specifically, the model selected the vector corresponding to the first column [CLS] of the token sequence as the feature vector of annotated texts. As a special token in the BERT model, [CLS] is used to mark the beginning of the input sequence, and its corresponding vector is trained to capture the global information of the entire input text. Therefore, it is very suitable to use it as a representation of the overall class information of annotated texts.

Figure 3 shows the text channel structure diagram. Let $y = \{q_1, ..., q_b\}$ be the annotated texts, i.e., the input of the text channel, and b be the maximum word sequence length. The input sequence of annotated texts was represented as follows:

$$Z = \{[cls], q_1, ..., q_b, [sep]\} \tag{5}$$

Let $TF(\ )$ be a layer of Transformer network, and Z be the input sequence with added marks of [*cls*] and [*sep*]. The equation of the output of the i-th layer Transformer in BERT was as follows:

$$G^u = \begin{cases} TF(Z), u = 0 \\ TF(G^{u-1}), 0 < u \le e \end{cases} \tag{6}$$

Let $g^e_{[cls]}$ be the [*cls*] column vector output by the last layer of the Transformer. After the output sequence of the last layer of the Transformer was expanded, there were:

$$G^e = \left\{ g^e_{[cls]}, g^e_1, g^e_2, ..., g^e_b, g^e_{[sep]} \right\}, e = 11 \tag{7}$$

Let Q^y_2 be the weight transposition of the dense layer, and n_2 be the offset. The following equation was used to obtain the output of the [*cls*] column vector of G^e after passing through the dense layer:

$$F = \text{ReLU}(Q^y_2 g^e[cls] + n_2) \tag{8}$$

Finally, the feature vectors of audio and annotated texts were fused at the feature level, obtaining a splicing vector. Then the model performed decision-level average weighting, and fusion classification on the feature vectors of audios and annotated texts and the splicing vector, which obtained the final class of multimedia audio learning resources, aiming to fully utilize the complementarity of multi-modality information to improve the accuracy and robustness of classification.

Let S be the feature vector of the audio channel output by the CNN model, Y be the feature vector of the annotated text channel output by the BERT model, P be the splicing feature obtained through feature splicing of S and Y at the feature level, and P_u be the u-th feature element. The equation of the probability distribution ε_u of multimedia audio learning resources in various classes was as follows:

$$\varepsilon_u = \text{softmax}(O) = \frac{\exp(P_u)}{\sum_{k=1}^{Y} \exp(P_k)};\ 1 \le k \le u \tag{9}$$

Let β_u be the class probability distribution of the audio channel calculated at the decision level, α_u be the class probability distribution of the annotated text channel, and *WE* be the average weighting operation. The equation of the final probability distribution results of multimedia audio learning resources in various classes was as follows:

$$\beta_u = \text{softmax}(S) \tag{10}$$

$$\alpha_u = \text{softmax}(Y) \tag{11}$$

$$OUT = WE(\beta_u, \alpha_u) \tag{12}$$

Let β_u, α_u, and ε_u be the class probability distributions of S, Y, and P, respectively. Average weighting was performed for the three distributions by combining the fusion structure at the feature and decision levels, which obtained the probability distribution results of classifying the multimedia audio learning resources using the constructed model.

$$P = CON(Y, S) \tag{13}$$

$$\beta_u = SM(S) \tag{14}$$

$$\alpha_u = \text{softmax}(Y) \tag{15}$$

$$\varepsilon_u = \text{softmax}(P) \tag{16}$$

$$OUT = WE(\beta_u, \alpha_u, \varepsilon_u) \tag{17}$$

3 RETRIEVING MULTIMEDIA AUDIO LEARNING RESOURCES BASED ON SELF-SIMILARITY MATRIX FILTERING

Audio fingerprinting is the feature representation of an audio signal that captures key features of audio and provides an effective method to compare and retrieve audio. An audio fingerprint is composed of a series of sub-fingerprints arranged in chronological order. Therefore, the temporal relationship between sub-fingerprints can be regarded as an important feature of audio. It is feasible to construct an index based on the similarity relationship of audio fingerprints to retrieve audio.

This feasibility is reflected in two aspects. On the one hand, audio signals are a kind of time series data, and their temporal structure is an important feature of audio content. The temporal relationship between sub-fingerprints captures this time structure, which allows us to consider the dynamic changes of audio, providing richer and more accurate audio information. The similarity relationship between sub-fingerprints in different audio fingerprints is generally different, which helps distinguish different audios by comparing the similarity relationship, thereby increasing the accuracy of audio retrieval. On the other hand, as audio data increases, the index can be continuously updated to maintain its coverage of the latest data. At the same time, the similarity relationship can be adjusted as necessary to make the index adapt to different retrieval needs.

In this study, a self-similarity matrix was used to represent the similarity relationship between fingerprint sequences. For the audio signal sequence $X = \{x_1, x_2, ..., x_b\}$ of multimedia audio learning resources, let a_{uk} be the similarity between x_u and x_k in the sequence, and then the self-similarity matrix was as follows:

$$AAL(X) = [a_{uk}]_{u,k=1,2,...,b} = \begin{bmatrix} a_{11} & a_{12} \cdots a_{1b} \\ a_{21} & a_{22} \cdots a_{2b} \\ \vdots & \vdots \quad \vdots \\ a_{b1} & a_{b2} \cdots a_{bb} \end{bmatrix} \tag{18}$$

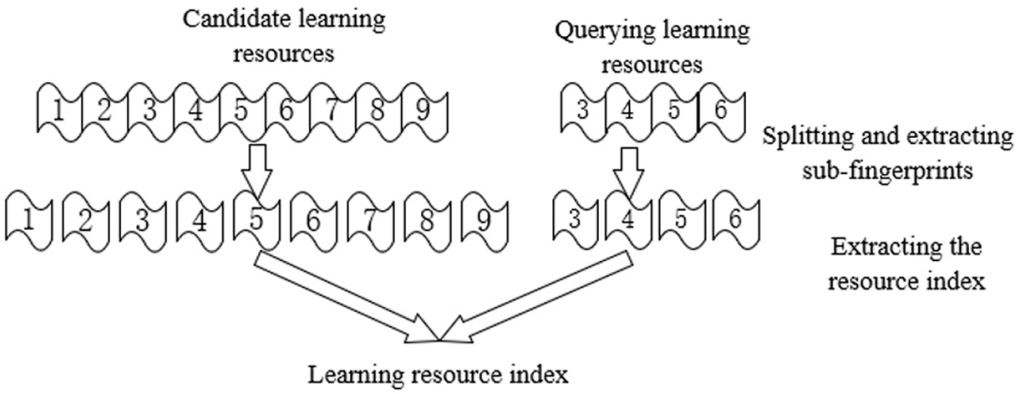

Fig. 4. Query and index extraction process of candidate learning resources

This study proposes to construct an index of multimedia audio learning resources based on a self-similarity matrix. Figure 4 shows a schematic diagram of the query and index extraction processes for candidate learning resources. Let $Z = \{z_1, z_2, ..., z_b\}$ be the audio fingerprint for querying multimedia audio learning resources, s_{uk} be the similarity between the audio sub-fingerprints z_u and z_k, and $S \in R^{b \times b}$ be the $b \times b$ floating-point matrix. The self-similarity matrix for Z was as follows:

$$S = AAL(Z) = [s_{uk}]u, k = 1,2,...,b \tag{19}$$

Let T be the candidate multimedia audio learning resources in the candidate set, and $N = AAL(T)$ be the self-similarity matrix of T, with $S,N \in R^{b\times b}$. The equation of the indicator to measure the similarity matching degree of indexes S and N based on the self-similarity matrix is given below. Let $S_u \in R^{l\times l}$ be the index of querying multimedia audio learning resources under the u-th aligned position; $N \in R^{b\times b}$ be the index of candidate multimedia audio learning resources, with $S = \{S_u\}_{u=1}^{5}$ and $l < b$; s_{uk} and n_{uk} be the values in the u-th row in the k-th column in matrices S and N, respectively. The equation to measure the mean absolute error between matrices was given as follows:

$$MAE(S,N) = \frac{\sum_{u=1}^{b}\sum_{k=1}^{b}\left|s_{uk} - n_{uk}\right|}{b^2} \tag{20}$$

The equation to measure the mean deviation of absolute errors between matrices was as follows:

$$MA = MADE(S,N) = \frac{\sum_{u=1}^{b}\sum_{k=1}^{b}\left|\left|s_{uk} - n_{uk}\right| - MAE\right|}{b^2} \tag{21}$$

The index matching of multimedia audio learning resources was considered a gradual matching process of S_u along the diagonal because the local relationship of the index was contained in the diagonal matrix of the self-similarity matrix. After going through all steps, the minimum value MAE_{MIN} of the mean absolute error was used to measure the matching degree of index and record the mean deviation of the corresponding absolute errors and the matching position m_2-PP at this time. Let $N_{u,k} \in E^{k\times k}$ be the sub-matrix $[n_{zt}]_{u\le z,t\le u+k}$ of N, and then the corresponding equation is as follows:

$$MAE_{MIN} = \underset{1\le j\le b-l+1}{MIN}(MAE(S_u, N_{j,l})) \tag{22}$$

$$m_2 - PP = \underset{1\le j\le b-l+1}{\operatorname{argmin}}(MAE(S_u, N_{j,l})) \tag{23}$$

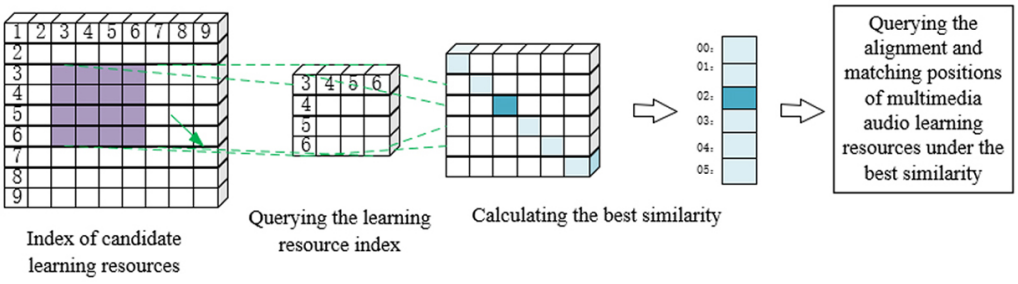

Fig. 5. Query and index matching process of candidate learning resources

Figure 5 shows the query and index matching processes for candidate learning resources. When querying each alignment position u of multimedia audio learning resources S, a group of calculation results were generated. The group of results with the lowest minimum value MAE_{MIN} of mean absolute error was selected as the final matching results of S, and the alignment position m_2-DQ of S was set to u. After going

through the filtered output candidate multimedia audio learning resource set U, the candidate multimedia audio learning resources whose mean absolute error and mean deviation of absolute errors were both less than the filtering threshold in the final matching results were stored in the candidate set U to complete the second filtering. The algorithm input included:

1. Querying the audio fingerprint $A = \{A^e\}_{u=1}^{5}$ of multimedia audio learning resources at different alignment positions, with $A = \{a, ..., a_v\}$;
2. The candidate multimedia audio learning resource set $U = \{E_u\}_{u=1}^{L}$, with E_u representing the audio fingerprint of the u-th candidate audio;
3. The reference multimedia audio learning resource library $FN = \{T_u\}_{u=1}^{B}$ used for index construction.

The second filtering was based on the similarity relationship between sub-fingerprints of multimedia audio learning resources and the candidate resource set U. It also included the query of alignment and matching positions of the resources under the best similarity. Therefore, the retrieval location information, which was output using the above methods, more accurately guided the matching task of the sub-fingerprint confirmation stage of the resources.

4 EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. Comparison of experimental results using different classification models

Models	Modalities	Accuracy (%)	Recall (%)	*F*1
SVM	Audios	0.78	0.78	0.78
RF	Audios	0.75	0.74	0.761
NBC	Annotated texts	0.63	0.63	0.63
BLR	Annotated texts	0.68	0.72	0.689
GBM	Annotated texts	0.66	0.66	0.68
DT	Annotated texts	0.62	0.61	0.61
MTL	Audios + annotated texts	0.81	0.77	0.798
The proposed model in this study	Audios + annotated texts	0.88	0.89	0.891

Table 2. F1-score comparison of processing single modality and multi-modal fusion using the proposed model

Modalities	Classes					*Macro Avg*
	Music Resources	Language Learning Resources	Lecture and Speech Resources	Film and TV Drama Resources	Audio Books and Radio Drama Resources	
Annotated texts	0.83	0.81	0.72	0.61	0.71	0.725
Audios	0.87	0.93	0.76	0.62	0.74	0.768
Multi-modal fusion	0.94	0.94	0.84	0.81	0.78	0.867

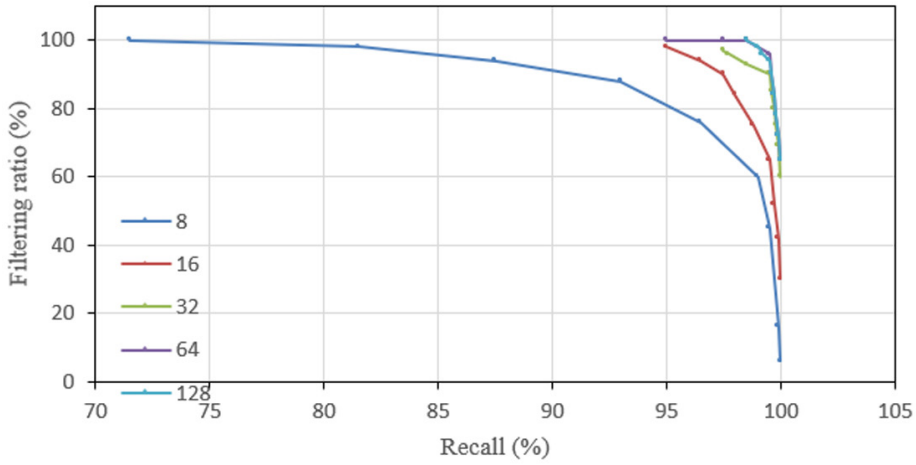

Fig. 6. Filtering performance of the retrieval method under different hash dimensions

It can be seen from Table 1 that various algorithms for classifying multimedia audio learning resources have different performances. For single-modality processing, that is, the processing of only audios or annotated texts, it can be observed that the performance of support vector machines (SVM) and random forests (RF) in processing audios is better than that of all algorithms in processing annotated texts. Among them, the performance of SVM in accuracy, recall, and F1-score is 0.78, while that of RF is 0.75, 0.74, and 0.761, respectively. For the processing of annotated texts, binary linear regression (BLR) has the best performance with 0.68 accuracy, 0.72 recall, and 0.689 F1-score, which is followed by gradient boosting machines (GBM) with 0.66 accuracy, 0.66 recall, and 0.68 F1-score, respectively. However, when both audio and annotated texts are processed simultaneously, for example, using the multi-modality processing method, the effect is significantly better than that of single-modality processing. The performance of multi-task learning (MTL) in accuracy, recall, and F1-score exceeds 0.77, while the proposed model in this study outperforms all other algorithms by exceeding 0.88 in all indicators. These results indicate that the multi-modality processing method that simultaneously processes audio and annotated texts is a better choice in the classification of multimedia audio learning resources, and the proposed model has the best performance among all the tested methods.

Table 2 shows in detail the macro average F1-score when using the single modality (annotation texts or audios) and multi-modal fusion processing methods for classifying different multimedia audio learning resources. When the model is used to process annotated texts only, the F1-score of the model is 0.83, 0.81, 0.72, 0.61, and 0.71, respectively, in the five resource classes, namely, music resources, language learning resources, lecture and speech resources, film and TV drama resources, audio books, and radio drama resources, with a macro average F1-score of 0.725. In contrast, the performance of the model improves when it is used to process audio only, and the F1-score is 0.87, 0.93, 0.76, 0.62, and 0.74 for the five classes, with a macro average F1-score of 0.768. However, when the model is used to simultaneously process audio and annotated texts, i.e., using the multi-modal fusion processing method, its performance significantly improves, and the F1-score increases to 0.94, 0.94, 0.84, 0.81, and 0.78, respectively, in the five classes, with a macro average F1-score also increasing to 0.867. In summary, the model proposed in this study performs better in classifying multimedia audio learning resources when simultaneously processing the multi-modal information (audio and annotated texts). Multi-modal fusion can

better capture the complementarity between audios and annotated texts, thereby improving the classification performance, especially in the classification task of multimedia audio learning resources in various classes.

Figure 6 evidently shows the variation in the relationship between recall and filtering ratio as the hash dimension increases. It can be observed that the recall improves in all stages as the hash dimension increases. It indicates that when the hash dimension increases, the retrieval algorithm can more accurately find relevant multimedia audio learning resources, i.e., the recall of the model increases. As the hash dimension increases, the decline rate of the filtering ratio also slows down in each stage, indicating that although the increase in the hash dimension improves recall, it also slows down the filtering efficiency, meaning that more candidate items need to be checked. It can be seen from the above observations that a balance between recall and filtering ratio should be found when choosing the hash dimension. An increase in the hash dimension improves recall but also reduces filtering efficiency. Therefore, to optimize retrieval performance, it is necessary to find a suitable hash dimension to ensure high recall while maintaining a high filtering ratio as much as possible.

Table 3. Comparison of experimental results using different retrieval methods

Methods	Query Duration (Second)	SNR(dB)	Database Size	Recall (%)	Accuracy (%)
The proposed method	6.1	0	10,000	98.2	97.8
LSH	6.2	0	500	61.5	–
NNS	5.1	0	10,000	81.1	–

According to Table 3, different retrieval methods can be compared and analyzed. It can be seen from the table that the proposed method in this study and locally sensitive hashing (LSH) have a similar query duration of about 6 seconds, while the nearest neighbor search (NNS) has a shorter query duration of 5.1 seconds. Although NNS has a relatively short query duration, its other performance indicators are not ideal. The signal-to-noise ratio (SNR) of all methods is 0, indicating a relatively clean testing environment with less noise. In terms of database size, both the proposed method and NNS process large-scale datasets, while LSH processes small-scale ones, which is one of the reasons for the poor performance of LSH because it may not be sufficient to process large-scale datasets. In terms of recall, the proposed method has a 98.2% recall, which is significantly superior to LSH and NNS, with a recalls of 81.1% and 61.5%, respectively. In addition, the proposed method also has excellent accuracy, reaching 97.8%, while the accuracy of LSH and NNS is not listed. Based on the above analysis, it can be seen that although the proposed method has a slightly longer query duration than NNS, it has excellent performance in key performance indicators such as the processing ability of large-scale datasets, recall, and accuracy. Therefore, the proposed method in this study has significant advantages for retrieving multimedia and audio learning resources.

According to Table 4, the retrieval performance of databases of different sizes can be compared and analyzed. The recall fluctuates slightly with the changes in database size. For small databases, such as the NBC speech database and LUMIAD, the recall is close to 98%, which is a very high value, indicating that almost all the relevant records are successfully retrieved. However, for large databases, such as MIREX DB and the Internet audio archive, the recall decreases but remains above

95%, which is still a relatively high proportion. The accuracy also varies under the influence of databases of different sizes, but the difference is not significant. All databases, regardless of their size, maintain an accuracy of over 95%, indicating that most of the retrieved records are relevant. The retrieval duration increases with the increase in database size, which is in line with expectations because the large amount of data means that more data needs to be retrieved and compared. The retrieval duration significantly increases from 29.4 milliseconds in the NBC speech database to 315.1 milliseconds in the Internet audio archive. Based on the above analysis, it can be seen that as the database size increases, the retrieval duration increases, while the recall may slightly decrease. However, the accuracy remains at a relatively high level in all cases, indicating that the retrieval method proposed in this study can still maintain high accuracy and recall when processing large-scale data. Although its retrieval duration may increase, the growth is still within an acceptable range.

Table 4. Comparison of retrieval performance under different data scales

Databases	Database Size (10,000)	Recall (%)	Accuracy (%)	Retrieval Duration (Millisecond)
NBC Speech Database	5	98.5	97.2	29.4
LUMIAD	10	98.1	97.0	65.2
MIREX DB	20	65.2	96.2	121.2
Internet Audio Archive	50	95.1	95.3	315.1

5 CONCLUSION

The study focused on the classification and retrieval of multimedia audio learning resources, especially integrating audio features with annotated text features to classify and retrieve audio.

In terms of classification, this study discussed the performance of various machine learning and deep learning methods under single-modality and multi-modality conditions. The experimental results showed that both audio and annotated texts obtained good classification effects. However, the multi-modal fusion method (MTL and the proposed model in this study) outperformed the single-modality method in terms of classification accuracy and recall.

In terms of retrieval, this study discussed how to retrieve by constructing the similarity relationship between audio fingerprints and comparing the filtering performance under different hash dimensions. The results showed that the retrieval performance gradually improved as the hash dimension increased, but the filter ratio slightly decreased. Among different retrieval methods, the proposed model outperformed LSH and NNS in terms of query duration, recall, and accuracy.

In addition, this study also conducted retrieval performance tests on databases of different sizes. The results showed that although the increase in database size led to an increase in retrieval duration, the proposed model in this study still maintained a high level of recall and accuracy.

In summary, this study demonstrated that the fusion of audio and annotated text features had a positive impact on the classification and retrieval of multimedia audio learning resources. The multi-modal fusion method showed high performance in both classification and retrieval. Moreover, the proposed model still maintained

high accuracy and recall when processing large-scale data. Although the retrieval duration may have increased, the increase was still within an acceptable range.

6 REFERENCES

[1] P. Wang, X. Wang, and X. Liu, "Selection of audio learning resources based on big data," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 6, pp. 23–38, 2022. https://doi.org/10.3991/ijet.v17i06.30013

[2] D. Stefani, S. Peroni, and L. Turchet, "A comparison of deep learning inference engines for embedded real-time audio classification," in *Proceedings of the International Conference on Digital Audio Effects*, DAFx, 2022, pp. 256–263.

[3] S. Ghosh, A. Seth, and S. Umesh, "Decorrelating feature spaces for learning general-purpose audio representations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1402–1414, 2022. https://doi.org/10.1109/JSTSP.2022.3202093

[4] K. K. Teh and H. D. Tran, "Open-set audio classification with limited training resources based on augmentation enhanced variational auto-encoder GAN with detection-classification joint training," *Interspeech*, pp. 4169–4173, 2021. https://doi.org/10.21437/Interspeech.2021-1142

[5] F. Carrión-Robles, V. Espinoza-Celi, and A. Vargas-Saritama, "The use of augmented reality through assemblr edu to inspire writing in an ecuadorian EFL distance program," *International Journal of Engineering Pedagogy*, vol. 13, no. 5, pp. 121–141, 2023. https://doi.org/10.3991/ijep.v13i5.38049

[6] S. Riahi, "Strengthening the teaching of soft skills in the pedagogical architecture of moroccan universities," *International Journal of Engineering Pedagogy*, vol. 12, no. 4, pp. 47–62, 2022. https://doi.org/10.3991/ijep.v12i4.22329

[7] H. Huang and C. T. Hsin, "Environmental literacy education and sustainable development in schools based on teaching effectiveness," *International Journal of Sustainable Development and Planning*, vol. 18, no. 5, pp. 1639–1648, 2023. https://doi.org/10.18280/ijsdp.180535

[8] S. Suryanti, W. Widodo, and Y. Yermiandhoko, "Gadget-based interactive multimedia on socio-scientific issues to improve elementary students' science literacy," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 1, pp. 56–69, 2021. https://doi.org/10.3991/ijim.v15i01.13675

[9] K. Q. Hussein, "A multimedia information time balance management in mobile cloud environment supported by case study," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 19, pp. 118–132, 2022. https://doi.org/10.3991/ijim.v16i19.33615

[10] A. Kumar, A. K. J. Saudagar, M. AlKhathami, B. Alsamani, M. H. A. Hasanat, M. B. Khan, A. Kumar, and K. U. Singh, "AIAVRT: 5.0 transformation in medical education with next generation AI- 3D animation and VR integrated computer graphics imagery," *Traitement du Signal*, vol. 39, no. 5, pp. 1823–1832, 2022. https://doi.org/10.18280/ts.390542

[11] Q. Liu, H. Chen, M.J.C. Crabbe, "Interactive study of multimedia and virtual technology in art education," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 1, pp. 80–93, 2021. https://doi.org/10.3991/ijet.v16i01.18227

[12] M. Budiarti, M. Ritonga, Rahmawati, Yasmadi, Julhadi, and Zulmuqim, "Padlet as a LMS platform in Arabic learning in higher education," *Ingénierie des Systèmes d'Information*, vol. 27, no. 4, pp. 659–664, 2022. https://doi.org/10.18280/isi.270417

[13] L. Baturina and A. Simakov, "Students' attitude towards e-learning in Russia after pandemic," *Education Science and Management*, vol. 1, no. 1, pp. 1–6, 2023. https://doi.org/10.56578/esm010101

[14] O. Habelko, N. Bozhko, I. Gavrysh, O. Khltobina, and Y. Necheporuk, "Characteristics of the influence of digital technologies on the system of learning a foreign language," *Ingénierie des Systèmes d'Information*, vol. 27, no. 5, pp. 835–841, 2022. https://doi.org/10.18280/isi.270518

[15] S. Fuada, "Development of educational kit for practical course in the topic of phase-shift RC oscillator," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 5, pp. 112–130, 2022. https://doi.org/10.3991/ijoe.v18i05.29131

[16] K. Krismadinata, U. Isni Kurnia, R. Mulya, and U. Verawardina, "The interactive multi media learning for power electronics course," *International Journal of Online and Biomedical Engineering*, vol. 18, no. 7, pp. 44–56, 2022. https://doi.org/10.3991/ijoe.v18i07.30029

[17] J. Rukayah, Daryanto, I. R. W. Atmojo, R. Ardiansyah, D. Y. Saputri, and M. Salimi, "Augmented reality media development in STEAM learning in elementary schools," *Ingénierie des Systèmes d'Information*, vol. 27, no. 3, pp. 463–471, 2022. https://doi.org/10.18280/isi.270313

[18] M. Mohaimenuzzaman, C. Bergmeir, and B. Meyer, "Pruning vs XNOR-net: A comprehensive study of deep learning for audio classification on edge-devices," *IEEE Access*, vol. 10, pp. 6696–6707, 2022. https://doi.org/10.1109/ACCESS.2022.3140807

[19] J. Pu, Y. Panagakis, and M. Pantic, "Learning separable time-frequency filterbanks for audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 3000–3004. https://doi.org/10.1109/ICASSP39728.2021.9414916

[20] J. Liu, J. Yang, W. W. Yang, and X. F. Kuang, "Research on the design of primary school English learning resources based on cognitive model," in *International Symposium on Educational Technology (ISET)*, Tokai, Nagoya, Japan, 2021, pp. 176–181. https://doi.org/10.1109/ISET52350.2021.00044

[21] A. L. Karn, J. L. Webber, A. Mehbodniya, D. Stalin David, B. Subramaniam, R. Rangasamy, and S. Sengan, "Evaluation and language training of multinational enterprises employees by deep learning in cloud manufacturing resources," in *International Conference on Innovations in Computer Science and Engineering*, 2022, pp. 369–380. https://doi.org/10.1007/978-981-19-7455-7_28

[22] X. Gong and Z. Li, "Improved video classification method based on non-parametric attention combined with self-supervision," in *Fourteenth International Conference on Digital Image Processing (ICDIP)*, vol. 12342, 2022, pp. 530–539. https://doi.org/10.1117/12.2643038

[23] W. Xie, Y. Li, Q. He, and W. Cao, "Few-shot class-incremental audio classification via discriminative prototype learning," *Expert Systems with Applications*, vol. 225, p. 120044, 2023. https://doi.org/10.1016/j.eswa.2023.120044

[24] A. S. Koepke, A. M. Oncescu, J. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, pp. 2675–2685, 2022. https://doi.org/10.1109/TMM.2022.3149712

[25] D. Takeuchi, Y. Ohishi, D. Niizumi, N. Harada, and K. Kashino, "Introducing auxiliary text query-modifier to content-based audio retrieval," *ArXiv Preprint*, 2022. https://doi.org/10.21437/Interspeech.2022-11428

[26] X. Hao, W. Zhang, D. Wu, F. Zhu, and B. Li, "Listen and look: Multi-modal aggregation and co-attention network for video-audio retrieval," in *IEEE International Conference on Multimedia and Expo (ICME)*, Taipei, Taiwan, 2022, pp. 1–6. https://doi.org/10.1109/ICME52920.2022.9859647

[27] Q. Zhang, J. Yang, X. Zhang, and T. Cao, "Generating adversarial examples in audio classification with generative adversarial network," in *7th International Conference on Image, Vision and Computing (ICIVC)*, Xi'an, China, 2022, pp. 848–853. https://doi.org/10.1109/ICIVC55077.2022.9886154

[28] M. Monjur and S. Nirjon, "An empirical analysis of perforated audio classification," in *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, 2022, pp. 25–30. https://doi.org/10.1145/3539490.3539602

[29] A. Koh and C. E. Siong, "Language-based audio retrieval with converging tied layers and contrastive loss," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Chiang Mai, Thailand, 2022, pp. 1644–1648. https://doi.org/10.23919/APSIPAASC55919.2022.9979840

[30] S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, "Neural audio fingerprint for high-specific audio retrieval based on contrastive learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 3025–3029. https://doi.org/10.1109/ICASSP39728.2021.9414337

7 AUTHOR

Wenwen Zhang currently holds a position as a Lecturer in the Music Department at Shijiazhuang Preschool Teachers College. She earned her degree from the Conservatory of Music at Hebei Normal University. Her research primarily focuses on musicology and music education. To date, she has published four academic papers, edited two textbooks, and authored a monograph (E-mail: zhangwenwen@hbafa.edu.cn; ORCID: https://orcid.org/0009-0003-2252-1034).