Audio-Text Retrieval Based on Contrastive Learning and Collaborative Attention Mechanism

Tao Hu
Central South University of Forestry and Technology
Xuyu Xiang (xyuxiang@csuft.edu.cn)
Central South University of Forestry and Technology
Jiaohua Qin
Central South University of Forestry and Technology
Yun Tan
Central South University of Forestry and Technology

Research Article

Keywords: Audio-text retrieval, audio augmentation, contrastive learning, collaborative attention mechanism

Posted Date: December 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2371994/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on August 2nd, 2023. See the published version at https://doi.org/10.1007/s00530-023-01144-4.

Abstract

Existing research on audio-text retrieval is limited by the size of the dataset and the structure of the network, making it difficult to learn the ideal featuresof audio and text resulting in low retrieval accuracy. In this paper, we construct an audio-text retrieval model based on contrastive learning and collaborative attention mechanism . We first reduce model overfitting by implementing audio augmentation strategies including adding Gaussian noise, adjusting the pitch and changing the time shift.Additionally, we design a co-attentive mechanism module that the audio data and text data guide each other in feature learning, effectively capturing the connection between the audio modality and the text modality. Finally we apply the contrastive learning methods between the augmented audio data and the original audio, allowing the model to effectively learn a richer set of audio features. The retrieval accuracy of our proposed model is significantly improved on publicly available datasets AudioCaps and Clotho.

1. Introduction

In recent years, multimedia data, including image, text, video and audio, has flooded our daily lives and this media information has become the main form of understanding the world. With more and more multimedia data, the previous single-media retrieval methods can no longer meet people's needs, and how to effectively perform multimodal information retrieval has become a popular research topic. Audio-text retrieval technology has very broad application scenarios in the age of information technology. First of all, people expect to be able to use text to retrieve audio clips on search engines and social networking software such as Google and Instagram in the same way as text retrieval for images and text retrieval for news, which will greatly enrich people's daily lives. Audio-text retrieval technology can also be applied to speech recognition, audio auditing and more. At the same time, the maturity of audio-text retrieval technology will make it easier to retrieve large volumes of media databases and facilitate the effective management of multimedia databases. As audio and text are important components of multimedia data, our aim is to construct a novel audio-text retrieval model.

Over the past decade, most research on audio retrieval has been devoted to content-based retrieval, that is, finding audios similar to the query audio from a reference audio database; however, content-based audio retrieval is limited by the structure of audio events, and audio retrieval tasks often perform poorly if the audio events are unstructured. We propose an audio-text retrieval framework that can query audio through detailed free-form natural language. For example, if people want to search for an audio clip of a dog barking after a thunderclap, they can use a text description like "After a thunderclap, the dog in the yard barked", which would have a chronological sequence of audio events, rather than a text description like "After the dog in the yard barked, the thunder rumbled". Audio retrieval with free-form text queries facilitates more flexible and accurate audio retrieval tasks as shown in Figure 1.

Current cross-media retrieval methods based on deep learning typically include binary representation-based learning method and real value representation-based method. The binary representation-based learning method projects cross-modal data into a common Hamming space and uses hash codes for

retrieval. For example, Jiang et al. [1] learns discrete hash codes through a cross-modal deep hashing algorithm to improve retrieval performance. To further improve the retrieval accuracy, Li et al. [2]proposed self-supervised adversarial hashing, which not only reduces a significant amount of time compared to the cross-modal deep hashing algorithm, but also learns richer supervised information. Wu et al. [3] proposed cycle-consistent deep generative hashing, which maximizes the relationship between the learned hash codes and each input and output, effectively compressing the data and maximizing the retention of its own information and the relationship between different modal samples. The binary representation-based learning method is highly efficient, with less storage space, and focuses on modal differences caused by modal feature heterogeneity, but it is difficult to solve the "semantic gap" problem and is not applicable in audio-text retrieval tasks.

The real value representation-based method has good semantic differentiation capabilities and can reduce the "semantic gap" to a large extent. Yu et al.[4] proposed a two-branch deep neural network learning architecture for audio modality and text modality. Lou et al. [5] evaluated the impact of various aggregation methods including mean pooling, max pooling, NetVLAD and NetRVLAD on cross-modal alignment for better audio-text alignment. Liu et al.[6]designed an omni-perception pre-trainer containing single encoders for audio and text modalities, and cross-modal encoders between the two modalities, to learn rich multimodal representations of audio and text. Manco et al. [7]proposed a framework for music contrastive audio-language learning. This is a dual-encoder architecture that learns cross-modal alignment between modalities and produces multimodal embeddings. Although these existing methods effectively focus on the relationship between audio and text modalities, they are limited by the size of the dataset to the extent that the modal features learned are often limited. Moreover, existing research on audio-text retrieval has focused too much on cross-modal connections, instead neglecting feature learning from a single modality. We address the problem of insufficient number of samples in existing audio-text retrieval by expanding the number of audio samples through audio augmentation. We propose a method to compare the augmented audio with the original audio to learn rich audio features, and we design a network structure based on collaborative attention mechanism to capture the close relationship between different modal data. We design an end-to-end network model that combines audio augmentation, multimodal feature extraction, contrastive learning, collaborative attention mechanism and common embedding space learning to improve the accuracy of mutual retrieval between audio and text as shown in Figure2. Our contributions in this paper are summarized as follows:

- Introducing audio augmentation and applying contrastive learning. In the audio-text retrieval, the introduction of audio augmentation not only solves the problem of insufficient sample data, but we also apply contrastive learning between the augmented audio and the original audio, and this self-supervision within the same modality effectively learns a richer set of features. We also evaluate the impact of different audio augmentation methods on audio-text retrieval, which provides more references for future application of our method to other cross-media tasks including audio.
- The collaborative attention module helps to learn more effective modal features. We use audio modality to assist in learning text features, and text modality to assist in learning audio features. The

collaborative attention module effectively captures the close relationship between audio and text, further improving retrieval accuracy.

- Comprehensive experiments show that our approach achieves excellent performance in audio-text retrieval.

2. Related Work

In this section, we briefly introduce audio-text retrieval, audio augmentation, and contrastive learning.

2.1 Audio-text Retrieval

The audio-text retrieval, as the name suggests, is to query the corresponding audio information through text, and query the corresponding text information through audio. Different media have inconsistent distributions and feature representations, and the main task in audio-text retrieval is to bridge the "heterogeneity gap" between the two. The current mainstream approach is to learn the feature representations of both modalities in a common embedding space, and the similarity between the two modalities is measured by the cosine similarity. How to learn the effective feature representations of both modalities has become a competing goal. Recently, Won et al. [8][9] successfully introduced multimodal metric learning for tag-based music retrieval, and focused on automatic retrieval of matching music for text-based stories. Zhang et al.[10] proposed a cross-modal audio-text retrieval method using an interactive learning convolutional autoencoder (CAE) to obtain shared features of audio and text patterns through interactive learning of CAE, which is then sent to a modal classifier to identify modal information for audio-text retrieval. Mei et al.[11]proposed an audio captioning system with an encoder-decoder architecture that uses transfer learning to alleviate problems caused by data scarcity, in addition to incorporating evaluation metrics into the optimization of reinforcement learning models. The audio-text retrieval is still at a preliminary stage of research compared to other inter-modal retrieval tasks, and more work is being done to establish suitable benchmarks. Kuzminykh et al. [12] investigates possible solutions for retrieving audio events based on natural language queries and evaluates the effectiveness and accuracy of multiple models. Koepke et al.[13] introduced new benchmarks for audio-text retrieval and used these to establish a baseline for cross-modal audio retrieval, demonstrating the benefits of pre-training for different audio tasks. In our work, we introduce audio augmentation from the perspective of limited audio-text data, drawing on image augmentation methods commonly used in image-related tasks. Moreover, we learn more complete audio features through contrastive learning, and we also design a collaborative attention-based network structure to further improve retrieval accuracy in terms of baseline metrics.

2.2 Audio augmentation

In the field of audio, the goal of audio augmentation is to enhance low-quality audio signals to improve quality and intelligibility. Audio augmentation methods are widely used for tasks such as speech recognition, speech separation and speech coding. The earliest methods of audio augmentation were

traditional statistical signal processing-based methods[14][15][16][17][18][19]. However, traditional methods cannot handle complex and irregular noise, and deep learning approaches can be more flexible to address these challenges. The most commonly used structure is the feedforward fully-connected neural network (FFNN) [20][21], in addition to the use of CNN[22][23], LSTM [24][25], BiLSTM[26][27]and others. Although there are many existing audio augmentation methods, it is a challenge to choose the appropriate augmentation method to be applied to audio-text retrieval. Unlike speech separation and speech synthesis, the purpose of applying audio augmentation is to expand the number of samples in audio-text retrieval. In reference image retrieval, images often need to be cropped and flipped, and the augmented audio should not be too different from the original audio. Therefore, we choose to use three audio augmentation methods: add Gaussian noise to the samples, pitch shift the sound up or down without changing the tempo and shift the samples forwards or backwards in this work.

2.3 Contrastive Learning

Contrastive learning [28][29][30][31] is a self-supervised learning method that learns general features of a dataset by allowing the model to learn which data are similar or dissimilar. Due to the excellent performance of contrastive learning on multimodal tasks, contrastive learning has become one of the popular methods to build multimodal retrieval models. Jia et al.[32][33][34][35] used contrastive learning to align representations of image and text achieving excellent feature representations and good results on their respective tasks. Contrastive learning has also been applied to NLP [36][37], including natural language understanding and machine translation tasks, where simple data augmentation has yielded results that approach or exceed SOTA. In cross-modal tasks [38][39], cross-modality contrastive learning is widely adopted to imply the information of different modalities into a unified semantic space. In this paper we apply contrastive learning not only to different modalities, but also within the same modality.

3. Methods

3.1 Problem formulation

Suppose we have a dataset *O* corresponding to text and audio, where O_t and O_a denote text and audio data respectively. We assume that $\{t_i, a_i\}_{i=1}^{N}$ are *N* sets of one-to-one corresponding text and audio instances, $t_i \in R^{d_t}$ denotes the d_t-dimensional text feature vector in dataset O_t, and $a_j \in R^{d_a}$ denotes the d_a-dimensional audio feature vector in dataset O_a. For a set of text-audio pairs (t_i,a_j) the similarity between them can be measured by the cosine similarity, as shown in Eq. 1.

$$s_{ij} = \frac{\phi(t_i) \cdot \psi(a_j)}{\|\phi(t_i)\|_2 \|\psi(a_j)\|_2}$$

1

where $\varphi()$ and $\psi()$ denote the encoders for text and audio, respectively. We consider s_{ii} as a positive pair where the text matches the audio, and s_{ij} as a negative pair where the text does not match the audio. Our

retrieval goal is to query the corresponding audio sample a_i by an arbitrary text sample t_i, and we also consider the opposite retrieval task, querying the corresponding text sample t_i by an arbitrary audio sample a_i.

3.2 Models

Previous experiments have shown that pre-trained models can achieve excellent results on cross-media retrieval tasks, so we used pre-trained models for text and audio feature extraction.

Text encoder: In the field of natural language processing (NLP), BERT has achieved state-of-the-art results on lots of tasks, and we chose to use the pre-trained BERT as a text encoder, appending a "<cls>" tag at the beginning of each sentence for the final whole-sentence feature representation. After the Bert encoder, the text feature dimension is B×768, where B is the batch size. After that, the text features are passed through two fully connected layers and a ReLu activation function, one fully connected layer is (768, 2048) and the other is (2048, 1024), and the final text feature dimension is adjusted to B×1024.

Audio encoder: For the choice of audio encoder, we refer to the work of Ref.[40] to select the audio encoder ResNet-38 used in PANNs, discarding two of the linear layers and applying the average pooling and max pooling layers to aggregate the frequency dimensions along the feature map output from the last convolution block to obtain an audio feature dimension of B × 2048. As with the text encoder, we pass the audio features through two fully connected layers and a ReLu activation function, one fully connected layer of (2048, 2048) and the other fully connected layer of (2048, 1024), and the audio feature dimension is similarly adjusted to B × 1024.

3.3 Collaborative attention mechanism

The collaborative attention mechanism refers to the self-attentive mechanism in Transformer, where there are three inputs Q (a matrix of query sets), K (a matrix of key sets), and V (a matrix of value sets). Various ranges of dependencies within the sequence can be captured through the Q, K, V attention mechanism. The attention mechanism through Q, K, V can capture various ranges of dependencies within the sequence. The dot product of each Q and K is divided by $\sqrt{d}$, and the attention weight is obtained after softmax processing, and then multiplied by the corresponding V as shown in formula 2, where d is the dimension of Q and K.

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

2

However, as opposed to learning just one attentional convergence, learning multiple attentional convergences, stitching the outputs of multiple attentional convergences together and varying them through another linear projection that can be learned to produce the final output can often be achieved

better. This approach is called multi-head attention and a single attentional convergence can be represented as:

$$\text{head}_{\text{i}} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

3

The final output is as in Eq. 4.

$$\text{MultiHead}\,(Q,K,V) = \text{Concat}\,(\text{head}_1,\ldots,\text{head m})\,W^O$$

4

Where $W^Q, W^K, W^V \epsilon R^{\text{d}\times d_m}$ $W^O \epsilon R^{\text{m}\times d_m \times d}$ is the learni, g projection matrix and d_m is the dimension of the output matrix. The vectors of *Q*, *K*, *V* are the same in the self attention mechanism. In the co-attentive mechanism module for audio-text retrieval, we input the text feature vector and the audio feature vector into the audio attention module, replacing the values of *K*,*V* with the text feature vector. The output F_A of the final audio features and the output F_T of the text features are obtained through a multi-head attention structure and a fully connected feed-forward network.

$$F_A = MultiHead\,(Q_A, K_T, V_T)$$

5

$$F_T = MultiHead\,(Q_T, K_A, V_A) \quad (6)$$

3.4 Audio augmentation

In a large number of text-image retrieval tasks, image augmentation is often used to generate similar image data to improve the robustness and generalization ability of the model. We feel that audio augmentation can also improve our retrieval efficiency in text-audio retrieval, and we have used several common audio augmentation methods: add Gaussian noise to the samples, pitch shift and time shift. We apply each augmentation method to the retrieval task individually or combine multiple augmentations to the retrieval task. The selection strategy for the three augmentation methods is random augmentation, and we set the probability of audio augmentation for the original audio as 0.5.

Add Gaussian noise: Gaussian noise is added to the audio samples with a minimum amplitude set to 0.001 and a maximum amplitude set to 0.015.

Pitch shift: pitch moves the sound up or down without changing the tempo, minimum semitone is set to -4 and maximum semitone is set to 4.

Time shift: shift the samples forwards or backwards, with or without rollover, we set the minimum fraction of total sound length to -0.5 and the maximum fraction of total sound length to 0.5, scrolling beyond the first or last position reintroduces the samples.

After three audio augmentation strategies, Mel spectrums of the original audio and augmented audios are shown in Fig. 3. From an intuitive point of view it is difficult to detect the subtle differences between them, but it is this subtle difference, which is not visible to the naked eye, that can obtain different audio features through the audio encoder. Although the audio encoder has captured most of the features of the original audio, there are still some details that are overlooked, and by contrastive learning between the original and the augmented audio, we can get richer features of the audio.

3.5 Contrastive learning

In the audio-text retrieval, we expand the number of audios by audio augmentation. We considered the original audio and the matched text as positive sample pairs, the augmented audio and text as positive sample pairs as well, and the other mismatched ones as negative sample pairs. We want the similarity between the positive samples to be as large as possible and the similarity between the negative samples to be as small as possible. Chen [41] learns visual representations in self-supervised learning and proposes a softmax-based contrast loss NT-Xent, which can be expressed as:

$$L1 = -\frac{1}{B}\sum_{i=1}^{B} log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)}$$

7

where *s*(*i*,*i*) denotes positive sample pairs and *s*(*i*,*j*) denotes negative sample pairs. *B* is the batch size and τ is the temperature coefficient, set to 0.2 in the experiments. Our audio retrieval task is bidirectional, so the loss function is computed in a bidirectional manner expressed as:

$$L2 = -\frac{1}{B}\left(\sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)} + \sum_{i=1}^{B}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ji}/\tau)}\right)$$

8

Intra-modal contrastive learning: we introduce intra-modal contrastive learning of audio on AudioCpas and Clotho. We consider the original audio a_i and the augmented audio $\widehat{a}_i$ as a positive sample pair *a*(*i*,*i*), then *a*(*i*,*j*) as a negative sample pair. The intra-modal contrastive learning of audio loss is :

$$L3 = -\frac{1}{B}\sum_{i=1}^{B} log\frac{\exp(a_{ii}/\tau)}{\sum_{j=1}^{B}\exp(a_{ij}/\tau)}$$

9

The final loss function can be expressed as:

$$\mathcal{L} = \mathcal{L}2 + \mathcal{L}3$$

10

4. Experiments

In this section, we first compare the performance of our model for audio-text retrieval (Text->Audio) and text audio-retrieval (Audio->Text) on both AudioCaps and Clotho datasets. We then conduct ablation experiments for each module in our model.

4.1 Datasets

AudioCaps

AudioCaps is a dataset for generating natural language descriptions for any type of audio data. The dataset consists of 46K pairs of audio clips and text descriptions, where the audio is mainly sourced from Audioset. The length of each audio clip is approximately 10*s*. The training set contains 49274 audio clips, where each audio clip corresponds to a text description, the validation set contains 494 audio clips, and the test set contains 957 audio clips, where each audio clip corresponds to five different text descriptions in the validation and test sets.

Clotho

Clotho is an audio captioning dataset consisting of 4981 audio samples, and the duration of each audio segment is 15-30*s*. The dataset is divided into a training set, a test set and a validation set. In Clotho v2, there are 3,839 audio clips in the training set and 1,045 clips in the validation and test sets. Each audio clip corresponds to five different text descriptions and each paragraph of text is 8 to 20 words in length.

4.2 Implementation details

In this section, we use the retrieval metrics of R@K (higher is better), median (MedR) and mean (MeanR) ranking (lower is better) to evaluate the performance of our model in retrieval tasks. R@K denotes the percentage of correct results retrieved in the top-K results, MedR denotes the median of the first correct result retrieved, and MeanR denotes the median of the first median of correct results retrieved.

During our experiments, the batch size in our experiments is set to 32, num_wokers is set to 6, the learning rate is 0.2 and the epoch in our experiments is set to 50 on AudioCpas. The batch size is set to 24, num_wokers is set to 8, the learning rate is 0.2 and epoch was set to 50 on Clotho.

4.3 Results

Our audio-text retrieval model is retrieved on AudioCpas and Clotho. We extract audio features using the pre-trained audio model ResNet38 in PANNs and text features via the pre-trained Bert model in HuggingFace, aligning the feature vectors of both modalities to 1024 latitude via a fully connected layer. We pass the audio and text feature vectors through the collaborative attentive module, using cross-modal contrastive learning to align features between the two modalities, and learning more effective single-modal features through inter-modal contrastive learning. We fine-tune the pre-trained audio and text encoders on the training set and select the model with the best combined performance on all retrieval metrics on the validation set to be applied to the test set. We perform two retrieval tasks, including audio

retrieval by text and text retrieval by audio. We compare our model with current state-of-the-art audio-text retrieval models and the results are shown in Table 1 and Table 2.

Table 1
Models for audio-text retrieval on AudioCaps

Model	**Text–>Audio**					
	R1↑	**R5↑**	**R10↑**	**R50↑**	**MedR↓**	**MeanR↓**
CNN14 + NetRVLAD[5]*	29.3 ± 0.3	65.2 ± 0.5	79.3 ± 1.0	/	3.0 ± 0.0	/
CE[13]	23.6 ± 0.6	56.2 ± 0.5	71.4 ± 0.5	92.3 ± 1.5	4.0 ± 0.0	18.3 ± 3.0
MOEE[13]	23.0 ± 0.7	55.7 ± 0.3	71.0 ± 1.2	93.0 ± 0.3	4.0 ± 0.0	16.3 ± 0.5
Ours	**33.4 ± 0.4**	**68.8 ± 0.1**	**81.9 ± 0.3**	**96.8 ± 0.2**	**3.0 ± 0.0**	**10.0 ± 0.3**
Model	Audio–>Text					
	R1↑	R5↑	R10↑	R50↑	MedR↓	MeanR↓
CNN14 + NetRVLAD[5]*	33.3 ± 0.5	67.6 ± 0.5	80.6 ± 0.8	/	3.0 ± 0.0	/
CE[13]	27.6 ± 1.0	60.5 ± 0.7	74.7 ± 0.8	94.2 ± 0.4	4.0 ± 0.0	14.7 ± 1.4
MOEE[13]	26.6 ± 0.7	59.3 ± 1.4	73.5 ± 1.1	94.0 ± 0.5	4.0 ± 0.0	15.6 ± 0.8
Ours	**42.3 ± 0.6**	**74.0 ± 0.7**	**85.3 ± 0.3**	**98.0 ± 0.2**	**2.0 ± 0.0**	**7.2 ± 0.3**

Table 2
Models for audio-text retrieval on Clotho

Model	**Text–>Audio**					
	R1↑	**R5↑**	**R10↑**	**R50↑**	**MedR↓**	**MeanR↓**
CNN14 + NetRVLAD[5]*	13.1 ± 0.2	33.1 ± 0.6	45.1 ± 0.2	/	13.0 ± 0.0	/
CE[13]	6.7 ± 0.4	21.6 ± 0.6	33.2 ± 0.3	69.8 ± 0.3	22.3 ± 0.6	58.3 ± 1.1
MOEE[13]	6.0 ± 0.1	20.8 ± 0.7	32.3 ± 0.3	68.5 ± 0.5	23.0 ± 0.0	60.2 ± 0.8
Ours	**12.7 ± 0.3**	**34.5 ± 0.7**	**47.1 ± 0.2**	**77.5 ± 0.4**	**12.0 ± 0.0**	**51.6 ± 1.3**
Model	Audio–>Text					
	R1↑	R5↑	R10↑	R50↑	MedR↓	MeanR↓
CNN14 + NetRVLAD[5]*	13.0 ± 0.2	32.9 ± 0.7	45.4 ± 0.8	/	13.0 ± 0.0	/
CE[13]	7.0 ± 0.3	22.7 ± 0.6	34.6 ± 0.5	67.9 ± 2.3	21.3 ± 0.6	72.6 ± 3.4
MOEE[13]	7.2 ± 0.5	22.1 ± 0.7	33.2 ± 1.1	67.4 ± 0.3	22.7 ± 0.6	71.8 ± 2.3
Ours	**14.3 ± 1.1**	**35.1 ± 1.0**	**48.1 ± 1.6**	**79.8 ± 0.6**	**11.3 ± 0.9**	**42.3 ± 2.2**

Note: The * indicates that the relevant source code is missing the results of this experiment are from the original paper, and / indicates that the metric is not given in the original paper.

Our work achieves superior retrieval results on the audio-text retrieval relative to previous work. On AudioCaps, our model improves R1 by 10%, R5 by 13%, R10 by 10%, R50 by 13%, MedR by 1% and MeanR by 6% on the task of text retrieval for audio relative to CE and MOEE. For the task of audio retrieval of text, R1 improved by 12%, R5 by 15%, R10 by 11%, R50 by 14%, MedR by 2% and MeanR by 7%. Our model has a combined improvement of over 3% on the text retrieval audio task and nearly 7% on the audio retrieval text task for the method of Ref.[5]. We believe that the main reason for the significantly higher improvement in the task of retrieving text by audio compared to retrieving audio by text is that the contrastive learning between audio modalities learns richer audio features, so that our "questions" are described more specifically in the retrieval process to the extent that our "answer" needs to be more precise.

On Clotho, our model improves R@1 by 6%, R@5 by 13%, R@10 by 14%, R50 by 8%, MedR and MeanR by nearly 10% compared to CE and MOEE on text retrieval audio task. On the audio retrieval text task,R@1 increases by 7%, R@5 increases by 13%, R@10 increases by 14%, R50 increases by 12%, and MedR and MeanR increase by close to 11%. Compared with the work of [5], our model reduces R@1 by 0.4% on the text retrieval audio task, and improves other indicators by about 2%. On the audio retrieval text task, R@1 increases by 1%, and other indicators increase by nearly 3%. Compared with AudioCaps, the Clotho dataset is more complicated to process, and the improvement of our model on Clotho is relatively reduced, but it still has a significant improvement compared to existing models. Then we evaluate the

experimental results of fine-tuning and freeze on AudioCaps and Clotho, as shown in Table 3 and Table 4.

Table 3
Experimental results of freeze and fine-tune for our model retrieval on AudioCpas

AudioCaps	**Text–>Audio**					
	R1↑	**R5↑**	**R10↑**	**R50↑**	**MedR↓**	**MeanR↓**
Freeze	19.7 ± 0.3	51.8 ± 0.3	68.2 ± 0.1	92.7 ± 0.3	5.0 ± 0.0	16.6 ± 0.2
Fine-tune	33.4 ± 0.4	68.8 ± 0.1	81.9 ± 0.3	96.8 ± 0.2	3.0 ± 0.0	10.0 ± 0.3
AudioCaps	Audio->Text					
	R1↑	R5↑	R10↑	R50↑	MedR↓	MeanR↓
Freeze	23.6 ± 0.2	56.8 ± 0.9	72.0 ± 1.3	95.1 ± 0.3	4.0 ± 0.0	12.5 ± 0.2
Fine-tune	42.3 ± 0.6	74.0 ± 0.7	85.3 ± 0.3	98.0 ± 0.2	2.0 ± 0.0	7.2 ± 0.3

Table 4
Experimental results of freeze and fine-tune for our model retrieval on Clotho

Clotho	**Text–>Audio**					
	R1↑	**R5↑**	**R10↑**	**R50↑**	**MedR↓**	**MeanR↓**
Freeze	8.4 ± 0.1	25.7 ± 0.6	38.1 ± 0.6	72.0 ± 0.1	18.3 ± 0.5	56.2 ± 0.5
Fine-tune	12.7 ± 0.3	34.5 ± 0.7	47.1 ± 0.2	77.5 ± 0.4	12.0 ± 0.0	51.6 ± 1.3
Clotho	Audio->Text					
	R1↑	R5↑	R10↑	R50↑	MedR↓	MeanR↓
Freeze	10.2 ± 0.8	27.8 ± 1.4	39.7 ± 1.3	71.8 ± 0.9	17.0 ± 0.8	64.5 ± 3.3
Fine-tune	14.3 ± 1.1	35.1 ± 1.0	48.1 ± 1.6	79.8 ± 0.6	11.3 ± 0.9	42.3 ± 2.2

When fine-tune of the pre-trained audio and text encoders on the training sets of the AudioCpas and Clotho, the retrieval accuracy of our model is significantly improved. Using pre-trained models and fine-tuning them on downstream tasks can significantly improve task performance, and fine-tuning is widely used in the fields of computer vision and natural language processing.

4.4 Ablation experiments

In the ablation experiments, we follow the implementation details in 4.1. Since fine-tuning would substantially increase the training time of the model, in this section none of the pre-trained encoders in our experiments are fine-tuned on the training set. We sequentially evaluate the effects of audio augmentation, collaborative attentive mechanism and inter-modal contrastive learning on audio-text retrieval in comparison experiments. The baseline model we use is the model with all three components of audio augmentation, collaborative attentive mechanism and inter-modal contrastive learning removed from our model.

4.4.1 Effect of audio augmentation

In the field of deep learning, data augmentation has been an important tool to improve the performance of a task. In the field of computer vision, strategies for image augmentation are found everywhere. In audio-text retrieval, our introduction of audio augmentation not only expands the dataset, but also provides a solution for contrastive learning between audio modalities in our follow-up work. We add audio augmentation module to the baseline model. The audio augmentation methods include adding Gaussian noise, pitch shift and time shift, then we combine the three augmentation methods two by two, and finally the three augmentation methods are combined. We add each of the different audio augmentation methods to the baseline model. The impact of the different audio augmentation methods on the experiments are evaluated on AudioCpas, as shown in Table 5.

Table 5
Different audio augmentation methods for audio-text retrieval on AudioCpas

Augmentation	**Text–>Audio**				
	R1↑	**R5↑**	**R10↑**	**MedR↓**	**MeanR↓**
No augmentation	18.9 ± 0.5	50.4 ± 0.3	66.2 ± 0.2	5.0 ± 0.0	19.2 ± 0.5
Add Gaussian noise+	19.2 ± 0.2	51.3 ± 0.3	67.4 ± 0.6	5.0 ± 0.0	17.7 ± 0.6
Time shift+	19.1 ± 0.1	51.1 ± 0.2	66.8 ± 0.2	5.0 ± 0.0	19.1 ± 0.2
Pitch shift+	19.2 ± 0.3	51.2 ± 0.5	67.0 ± 0.2	5.0 ± 0.0	17.6 ± 0.1
Add Gaussian noise+ Pitch shift+	19.2 ± 0.2	51.5 ± 0.2	68.0 ± 0.1	5.0 ± 0.0	17.2 ± 0.2
Add Gaussian noise+ Time shift+	19.7 ± 0.4	51.5 ± 0.4	68.0 ± 0.2	5.0 ± 0.0	17.3 ± 0.2
Time shift+ Pitch shift+	19.1 ± 0.1	51.7 ± 0.2	67.6 ± 0.1	5.0 ± 0.0	17.4 ± 0.1
Mix+	18.6 ± 0.2	51.4 ± 0.1	68.3 ± 0.2	5.0 ± 0.0	16.7 ± 0.3
Augmentation	Audio->Text				
	R1↑	R5↑	R10↑	MedR↓	MeanR↓
No augmentation	20.3 ± 0.7	53.0 ± 0.8	69.6 ± 0.8	5.0 ± 0.0	16.9 ± 0.6
Add Gaussian noise+	21.3 ± 0.8	54.2 ± 0.5	70.7 ± 1.0	5.0 ± 0.0	14.3 ± 0.4
Time shift+	21.5 ± 0.1	54.0 ± 0.8	70.1 ± 0.7	5.0 ± 0.0	15.7 ± 0.1
Pitch shift+	20.9 ± 0.5	54.1 ± 0.6	70.1 ± 0.3	5.0 ± 0.0	15.4 ± 0.2
Add Gaussian noise+ Pitch shift+	20.9 ± 1.6	53.1 ± 1.4	70.4 ± 0.8	5.0 ± 0.0	14.9 ± 0.4
Add Gaussian noise+ Time shift+	21.6 ± 0.6	54.7 ± 0.8	71.4 ± 0.2	5.0 ± 0.0	14.2 ± 0.2
Time shift+ Pitch shift+	20.8 ± 0.3	53.0 ± 0.6	69.5 ± 0.8	5.0 ± 0.0	15.1 ± 0.4
Mix+	20.5 ± 0.8	53.3 ± 0.4	69.8 ± 0.9	5.0 ± 0.0	14.7 ± 0.4

We can observe that whether a single audio augmentation method is used or a combination of different audio augmentation methods, the performance improvement for the retrieval task is roughly the same for all methods, improving the retrieval metric by 0.5–2%, but the combination of adding Gaussian noise and time shift works best in relative terms. Mixing the three augmentation methods instead reduced the metric of R@1, and we believe that overly complex audio changes to the original audio were instead detrimental to its feature learning. It is worth noting that we found during the training process that the method of adjusting the audio pitch increases the time overhead substantially and does not lead to better enhancement. We think that adjusting the pitch is more altering to the original audio compared to the other two audio augmentation methods, which will change the frequency of the original audio itself

increasing the time overhead, so if the audio augmentation is used in related tasks by subsequent scholars, it can be discard this method.

4.4.2 Effect of the collaborative attention mechanism

The biggest challenge faced in cross-modal retrieval tasks is how to address the heterogeneity divide. Existing approaches have worked to reduce the disparity between different modalities. We introduce a collaborative attentive mechanism with reference to the attention mechanism in the audio-text retrieval, where information from the audio modality is used to guide feature learning in the text modality, and information from the text modality is used to guide feature extraction in the audio modality. We hope that this interaction of information between different modalities can appropriately reduce the variability between different modalities.

We add the collaborative mechanism to the baseline model. During the experiments, we use multiple heads of attention, with heads set to 2, 4 and 8, and dropout in the attention mechanism set to 0.2. We evaluate the effect of the collaborative attentive mechanism on AudioCpas and assess the variability of the different heads in collaborative attentive mechanism in the experiments. The boost of collaborative attentive mechanism relative to the audio-text retrieval task is approximately 0.5%, with the best results achieved when the heads equal 4, as shown in Table 6.

Table 6
Audio-text retrieval on AudioCpas for different sizes of heads in the collaborative attention mechanism

AudioCaps	**Text–>Audio**				
	R1↑	**R5↑**	**R10↑**	**MedR↓**	**MeanR↓**
No Co-attention	18.9 ± 0.5	50.4 ± 0.3	66.2 ± 0.2	5.0 ± 0.0	19.2 ± 0.5
Co-attention heads = 2	19.3 ± 0.1	50.8 ± 0.6	66.7 ± 0.3	5.0 ± 0.0	18.1 ± 0.2
Co-attention heads = 4	19.5 ± 0.2	51.1 ± 0.4	67.1 ± 0.6	5.0 ± 0.0	18.0 ± 0.5
Co-attention heads = 8	19.7 ± 0.3	51.2 ± 0.2	66.4 ± 0.3	5.0 ± 0.0	18.4 ± 0.2
AudioCaps	Audio->Text				
	R1↑	R5↑	R10↑	MedR↓	MeanR↓
No Co-attention	20.3 ± 0.7	53.0 ± 0.8	69.6 ± 0.8	5.0 ± 0.0	16.9 ± 0.6
Co-attention heads = 2	21.6 ± 1.0	53.3 ± 0.1	69.8 ± 0.5	5.0 ± 0.0	14.7 ± 0.5
Co-attention heads = 4	20.8 ± 0.9	54.2 ± 1.0	70.0 ± 0.2	4.6 ± 0.4	14.8 ± 0.6
Co-attention heads = 8	21.1 ± 0.7	54.4 ± 0.6	70.1 ± 0.3	5.0 ± 0.0	15.2 ± 0.4

4.4.3 Effect of the intra-modal contrastive learning

Referring to the experimental results in Table 3, we chose the audio augmentation method with the highest overall performance improvement, combining the strategy of adding Gaussian noise and time shift to augment the original audio. We first evaluate the effect of the intra-modal contrastive(IMC) learning module on the experiments on AudioCaps, as shown in Table 7.

Table 7
Effect of the intra-modal contrastive learning for audio-text retrieval on AudioCpas

AudioCaps	**Text–>Audio**				
	R1↑	**R5↑**	**R10↑**	**MedR↓**	**MeanR↓**
No IMC	18.9 ± 0.5	50.4 ± 0.3	66.2 ± 0.2	5.0 ± 0.0	19.2 ± 0.5
IMC	19.1 ± 0.7	51.3 ± 0.3	67.9 ± 0.5	5.0 ± 0.0	17.5 ± 0.1
AudioCaps	Audio->Text				
	R1↑	R5↑	R10↑	MedR↓	MeanR↓
No IMC	20.3 ± 0.7	53.0 ± 0.8	69.6 ± 0.8	5.0 ± 0.0	16.9 ± 0.6
IMC	23.0 ± 0.6	57.1 ± 1.1	71.1 ± 0.2	4.0 ± 0.0	14.2 ± 0.4

Contrastive learning within the audio modality has a relatively large improvement on the audio-text retrieval, particularly for the audio retrieval text task, where the maximum improvement can be up to 4%. Observing the complexity of the Clotho dataset, in addition to the comparison module within the audio modality, we also included a comparison module within the text modality, where we return two texts at a time from a given set of five texts for comparison learning. We conduct experiments on Clotho to evaluate its effectiveness, as shown in Table 8.

Table 8
Effect of the intra-modal contrastive learning for audio-text retrieval on Clotho

Clotho	**Text–>Audio**				
	R1↑	**R5↑**	**R10↑**	**MedR↓**	**MeanR↓**
No IMC	8.2 ± 0.2	25.2 ± 0.2	36.5 ± 0.1	20.0 ± 0.0	61.3 ± 0.5
Audio IMC	8.1 ± 0.1	25.1 ± 0.2	36.4 ± 0.1	19.7 ± 0.5	58.6 ± 1.4
Audio + Text IMC	7.9 ± 0.2	25.2 ± 0.1	37.0 ± 0.1	19.0 ± 0.0	57.8 ± 0.8
Clotho	Audio->Text				
	R1↑	R5↑	R10↑	MedR↓	MeanR↓
No IMC	9.8 ± 0.3	27.8 ± 0.4	38.7 ± 0.6	18.7 ± 0.9	68.8 ± 0.5
Audio IMC	9.7 ± 0.2	27.1 ± 0.8	38.9 ± 0.5	18.0 ± 0.0	67.7 ± 0.8
Audio + Text IMC	9.8 ± 0.1	28.0 ± 1.0	40.0 ± 0.6	17.3 ± 0.5	67.1 ± 1.6

We found that the method of contrastive learning in the audio modality has little improvement for retrieval tasks on the Clotho dataset other than MedR and MeanR. We carefully studied the data of the Clotho data set, as shown in Table 9. In the Clotho training set, each audio corresponds to five text descriptions, which are different in length and content. Coupled with the small number of samples in the Clotho dataset, it was difficult for us to effectively learn valid audio and text features, and even though we expanded the number of audios through audio enhancement, the fact that the text matched with them was not the same text each time greatly makes our retrieval more difficult. Our attempts to get a better feature representation using contrastive learning within the text modality, where five different texts learn from each other. In the task of text retrieval audio, R@10 increases 0.5%, and in the task of audio retrieval text R@10 improves 1.3%. We will follow up with further research on how to handle such difficult datasets.

5. Conclusion

In this paper, we introduced a new framework for audio-text retrieval in which we designed three modules for audio-text retrieval including audio augmentation, collaborative attention mechanism and intra-modal contrastive learning. We achieved excellent retrieval performance using pre-trained models and fine-tuning on datasets for audio-text retrieval. We evaluate the effectiveness of the three modules we introduce on the audio-text retrieval. In summary, our method provides advanced performance for audio-text retrieval, providing guidance for further research in audio-text retrieval.

References

1. Jiang Q Y, Li W J. Deep cross-modal hashing[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3232-3240.
2. Li C, Deng C, Li N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4242-4251.
3. Wu L, Wang Y, Shao L. Cycle-consistent deep generative hashing for cross-modal retrieval[J]. IEEE Transactions on Image Processing, 2018, 28(4): 1602-1612.
4. Yu Y, Tang S, Raposo F, et al. Deep cross-modal correlation learning for audio and lyrics in music retrieval[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2019, 15(1): 1-16.
5. Lou S, Xu X, Wu M, et al. Audio-Text Retrieval in Context[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 4793-4797.
6. Liu J, Zhu X, Liu F, et al. Opt: omni-perception pre-trainer for cross-modal understanding and generation[J]. https://arxiv.org/abs/2107.00249, 2021.
7. Manco I, Benetos E, Quinton E, et al. Contrastive audio-language learning for music[J]. https://arxiv.org/abs/2208.12208, 2022.
8. Won M, Oramas S, Nieto O, et al. Multimodal metric learning for tag-based music retrieval[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 591-595.
9. Won M, Salamon J, Bryan N J, et al. Emotion Embedding Spaces for Matching Music to Stories[J]. https://arxiv.org/abs/2111.13468, 2021.
10. Zhang, Hongli. "Voice keyword retrieval method using attention mechanism and multimodal information fusion." *Scientific Programming* 2021 (2021). Ma J, Gu X. Scene image retrieval with siamese spatial attention pooling[J]. Neurocomputing, 2020, 412: 252-261.
11. Mei X, Huang Q, Liu X, et al. An encoder-decoder based audio captioning system with transfer and reinforcement learning[J]. https://arxiv.org/abs/2108.02752, 2021.
12. Kuzminykh I, Shevchuk D, Shiaeles S, et al. Audio interval retrieval using convolutional neural networks[M]//Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer, Cham, 2020: 229-240.
13. Koepke A S, Oncescu A M, Henriques J, et al. Audio retrieval with natural language queries: A benchmark study[J]. IEEE Transactions on Multimedia, 2022.
14. Abel A, Hussain A. Novel two-stage audiovisual speech filtering in noisy environments[J]. Cognitive Computation, 2014, 6(2): 200-217.
15. Almajai I, Milner B. Visually derived wiener filters for speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(6): 1642-1651.
16. Khan M S, Naqvi S M, Wang W, et al. Video-aided model-based source separation in real reverberant rooms[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(9): 1900-1912.

17. Liang Y, Naqvi S M, Chambers J A. Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment[J]. EURASIP journal on Advances in Signal Processing, 2012, 2012(1): 1-16.
18. Maganti H K, Gatica-Perez D, McCowan I. Speech enhancement and recognition in meetings with an audio–visual sensor array[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(8): 2257-2269.
19. Rivet B, Girin L, Jutten C. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures[J]. IEEE transactions on audio, speech, and language processing, 2006, 15(1): 96-108.
20. Sadeghi M, Alameda-Pineda X. Mixture of inference networks for VAE-based audio-visual speech enhancement[J]. IEEE Transactions on Signal Processing, 2021, 69: 1899-1909.
21. Sadeghi M, Alameda-Pineda X. Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7534-7538.
22. Ideli E. Audio-visual speech processing using deep learning techniques[D]. Applied Sciences: School of Engineering Science, 2019.
23. Ideli E, Sharpe B, Bajić I V, et al. Visually assisted time-domain speech enhancement[C]//2019 IEEE global conference on signal and information processing (GlobalSIP). IEEE, 2019: 1-5.
24. Adeel A, Ahmad J, Larijani H, et al. A novel real-time, lightweight chaotic-encryption scheme for next-generation audio-visual hearing aids[J]. Cognitive Computation, 2020, 12(3): 589-601.
25. Adeel A, Gogate M, Hussain A. Towards next-generation lipreading driven hearing-aids: A preliminary prototype demo[C]//Proceedings of the International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017), Stockholm, Sweden. 2017, 19.
26. Afouras T, Chung J S, Zisserman A. My lips are concealed: Audio-visual speech enhancement through obstructions[J]. https://arxiv.org/abs/1907.04975, 2019.
27. Arriandiaga A, Morrone G, Pasa L, et al. Audio-visual target speaker enhancement on multi-talker environment using event-driven cameras[C]//2021 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2021: 1-5.
28. Wu Z, Xiong Y, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3733-3742.
29. Ye M, Zhang X, Yuen P C, et al. Unsupervised embedding learning via invariant and spreading instance feature[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6210-6219.
30. Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. https://arxiv.org/abs/1807.03748, 2018.
31. Tian Y, Krishnan D, Isola P. Contrastive multiview coding[C]//European conference on computer vision. Springer, Cham, 2020: 776-794.

32. Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//International Conference on Machine Learning. PMLR, 2021: 4904-4916.
33. Li J, Selvaraju R, Gotmare A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in neural information processing systems, 2021, 34: 9694-9705.
34. Wang W, Bao H, Dong L, et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts[J]. https://arxiv.org/abs/2111.02358, 2021.
35. Shen D, Zheng M, Shen Y, et al. A simple but tough-to-beat data augmentation approach for natural language understanding and generation[J]. arXiv preprint arXiv:2009.13818, 2020.
36. Fang H, Wang S, Zhou M, et al. Cert: Contrastive self-supervised learning for language understanding[J]. https://arxiv.org/abs/2005.12766, 2020.
37. Wu X, Gao C, Zang L, et al. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding[J]. https://arxiv.org/abs/2109.04380, 2021.
38. Li W, Gao C, Niu G, et al. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning[J]. https://arxiv.org/abs/2012.15409, 2020.
39. Zhang H, Koh J Y, Baldridge J, et al. Cross-modal contrastive learning for text-to-image generation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 833-842.
40. Mei X, Liu X, Sun J, et al. On Metric Learning for Audio-Text Cross-Modal Retrieval[J]. https://arxiv.org/abs/2203.15537, 2022.
41. Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

Table

Table 9 is available in the Supplementary Files section.

Figures

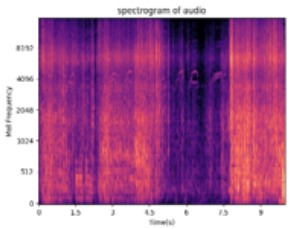
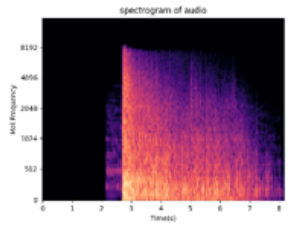
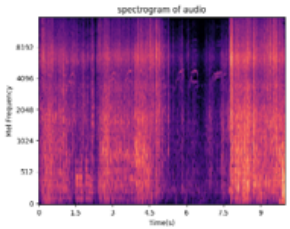
	Toilet flushing and water running down the drain.
	Thunder sounds loudly nearby.
	Several birds tweet and flap wings while a man whispers.

Figure 1

Matching natural language text to audio

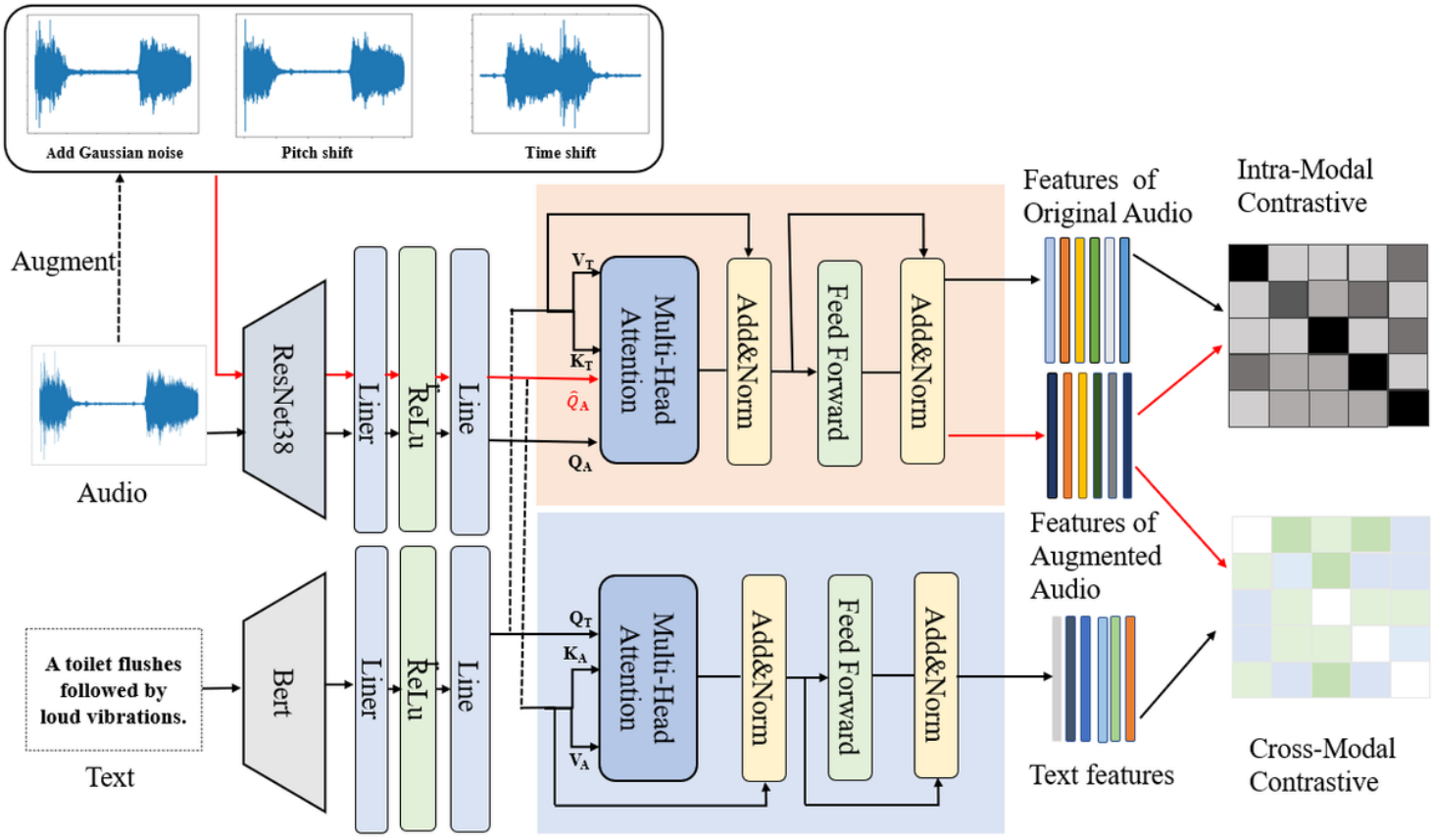


Figure 2

Overview of the audio-text retrieval model

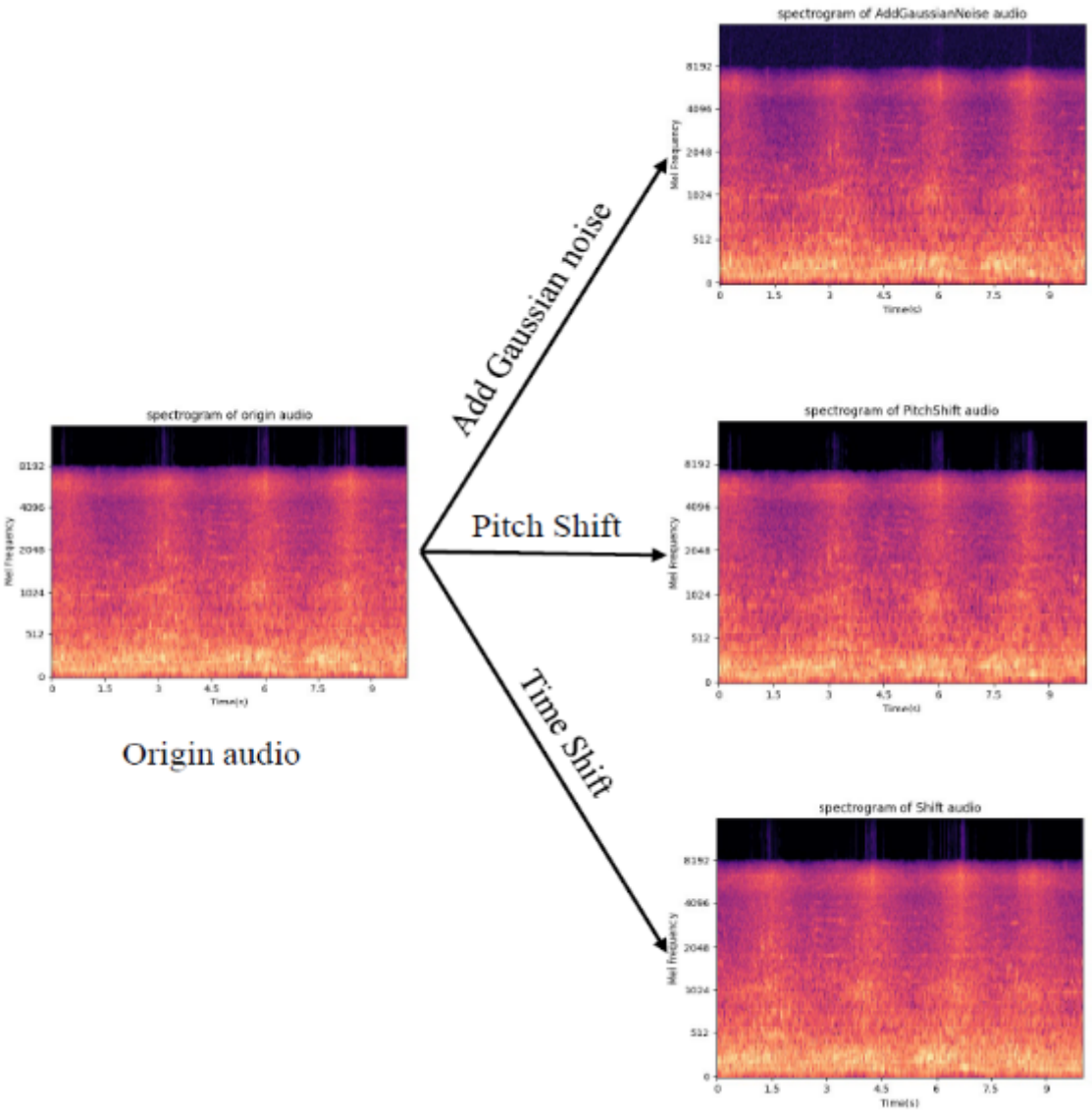


Figure 3

Mel spectrums of the original audio and augmented audios

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table9.docx