1-D convolutional neural network-based crosstalk quality assessment

Shuxin Yang

Department of Internet of Things Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China

yangshuxin@bupt.edu.cn

Abstract. Due to the current state of the declining field of crosstalk, many studies are engaged in investigating causes and devising strategies for its improvement. Nonetheless, there are often no quantitative measures available for assessing the quality of crosstalk. This study employed recordings made during traditional crosstalk performances to assess crosstalk quality based on the percentage of positive audience feedback sounds such as applause and laughter throughout the performance. This study compares Mel Frequency Cepstral Coefficients (MFCC) and Mel Filterbank Energies (MFE) audio feature extraction methods and compares different classification models by training a one-dimensional convolutional neural network model to explore the model that performs better in audio classification under low audio quality conditions, such as low sampling rate and small signal-to-noise ratio. In this study, the training data set has been divided into two different schemes. One is the sound of laughter, applause, singing and speaking, the other is the sound of speaking, singing and laughter&applause. In this study, the final performance of the different models, including accuracy and loss, is counted. The experimental results demonstrated that the models obtained when the MFE method is used and the audio classification is labeled as singing, applause and laughter, and performance speech training have better performance when the comic audio is classified.

Keywords: deep learning, audio recognition, 1D-CNN.

1. Introduction

Xiangsheng, known as "Cross Talk", is a traditional Chinese comedic performance that involves a humorous dialogue between two performers, utilizing puns, wordplay, and satire to entertain the audience. With a long history spanning over 100 years, Xiangsheng originated in Beijing and was initially performed on street corners by itinerant performers. Over time, it evolved into a more sophisticated art form and gained immense popularity in theatres and on television. Despite its enduring popularity, the art of cross talk has seen a decline due to various social and economic changes, including the rapid success of new media such as television [1]. To develop the art of cross talk, relying solely on social analysis is not enough, but also requires data support that can quantify the quality of cross talk. At present, there are few studies proposing models for evaluating the quality of cross talk, and most of them only have theoretical support without specific quantitative methods. Therefore, it is imperative to develop models that can objectively evaluate the quality of Xiangsheng performances to aid in their promotion and preservation.

Audio scene classification involves the identification of the acoustic environment of a given audio recording through the use of feature extraction techniques and classifier construction, which is a crucial technology in speech recognition. The commonly used feature extraction methods include log - mel features, Mel - Frequency Cepstral Coefficients (MFCC), and mel features. The purpose of feature extraction is to extract data that can replace the original audio signal. In recent years, 1-D convolutional neural networks (CNN) have been successfully applied in classification tasks, greatly improving accuracy, and have also been widely used in audio scene classification [2]. In recent years, one-dimensional convolutional neural networks (1D CNNs) have been widely applied to music classification tasks. By learning from a large dataset of musical samples, the 1D CNN model has shown remarkable accuracy in distinguishing between different genres of music such as jazz, rock, and others [3]. Furthermore, the ability of 1D CNNs to classify audio signals has practical applications in safety monitoring in production settings. For example, by analyzing the sound produced when a screw is hit, a 1D CNN can effectively determine whether the screw is loosened or not. This demonstrates the potential of 1D CNNs in addressing practical problems beyond the scope of traditional music classification tasks [4]. Similarly, one-dimensional convolutional neural networks have good performance in human voice classification and can even recognize the accent of a speaker based on speech features such as vowels and consonants [5]. Although advances in scientific technology have increased opportunities for obtaining high-quality audio datasets in recent years, neural networks -- especially one-dimensional convolutional neural networks -- experience rapid declines in learning accuracy when trained on low-quality data [3]. The neural network models trained on high-accuracy datasets are not well-suited for processing low-accuracy datasets. Considering that there are not many clips with relatively low sound quality in comic datasets, it is crucial to study the proposed adaptive audio classification model for low quality datasets in assessing comic quality.

This article employs machine learning methods to analyze audience feedback types in cross-talk performances. Based on existing research on convolutional neural networks, this paper proposes a new model to fill the research gap in the field of automatic evaluation of cross-talk quality. Specifically, this paper adjusts and optimizes the parameters of the original model, successfully improving the recognition accuracy of the neural network under low data quality. Experimental results show that the proposed model has high accuracy and reliability in evaluating cross-talk quality. Specifically, the feedback is divided into two types: positive and negative, which correspond to audience appreciation and dissatisfaction with the performer's presentation. By processing and analyzing these data, this study obtains an index that reflects the quality of cross-talk performances, which can be used to evaluate the performance's excellence. Moreover, singing is also an essential part of cross-talk performances. Therefore, speech signal processing techniques are employed to successfully identify the singing parts within the performances, which serve as an auxiliary indicator for evaluating the quality of cross-talk performances. Through these means, this article successfully constructs an effective model for evaluating the quality of cross-talk performances even through low-equality data set, providing significant support for promoting the inheritance and development of cross-talk culture.

2. Method

2.1. Dataset description and preprocessing

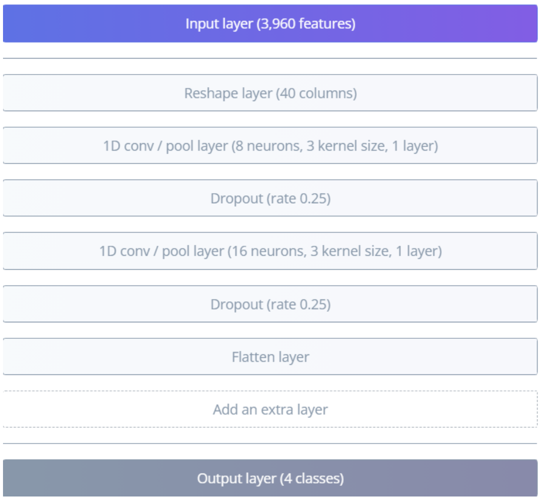
The data utilized in this paper was sourced from the renowned xiangsheng company, Deyun Society, spanning from 2000 to 2015. The audio dataset consists of five segments, each of which is 5-10 minutes in duration. The selection of Deyun Society's xiangsheng performances as the data source was motivated by various factors. Firstly, due to its long history and popularity, the audio recordings featuring performances by Deyun Society were well-preserved and exhibited a high degree of completeness, providing more comprehensive and accurate training data. Secondly, the audio quality of the dataset was superior, resulting in clear sound characteristics that facilitated further processing for comparative experiments. Additionally, the dataset includes xiangsheng works from different genres, time periods, and performers, representing a rich diversity of styles, techniques, and performance forms. This richness and diversity make the trained model more widely applicable, suitable for exploring practices in multiple fields.

Furthermore, the wide temporal span covered by the dataset encompasses a considerable period, from 2000 to 2015, and has historical and cultural significance. The temporal breadth of xiangsheng works included in the dataset also makes the trained model applicable to xiangsheng works from different eras, thereby exerting a broader and deeper impact on research and practice in this field. Therefore, this study utilizes the data to explore the development of xiangsheng speech synthesis technology and aims to contribute to the advancement of this field. This paper segmented the audio clips of these five crosstalk datasets into 2-second audio segments, with a total of approximately 1500 small segments. These small segments were manually classified into four categories, namely audience applause, audience laughter, crosstalk actor singing and speech sounds, thus obtaining training set 1. Concurrently, this study also added noise to the original data source and segmented it into 2-second training set 2 as well. The purpose of this approach is twofold: firstly, different sound features are more pronounced in small segments, facilitating model training; secondly, low signal-to-noise ratio datasets can enhance the performance of neural network models in noisy application scenarios.

2.2. 1D CNN model-based cross talk

In order to simplify the process of model deployment and application, this experiment employed EdgeImpulse platform for training neural network models and deployment. EdgeImpulse is an end-to-end open-source machine learning platform aimed to assists developers in building, optimizing, and deploying small machine learning models. The platform provides a range of tools and libraries for data acquisition, processing, and analysis, designing and training models, and deploying trained models to various embedded devices and microcontrollers. In the field of audio classification, EdgeImpulse platform combines advanced technologies such as deep convolutional neural networks (DCNN) and long short-term memory networks (LSTM) [6-8], enabling users to easily perform audio classification tasks. EdgeImpulse also provides an intuitive graphical user interface where users can drag and drop to implement an end-to-end data flow, including data acquisition, pre-processing, feature extraction, model training, and deployment. Additionally, EdgeImpulse offers APIs based on Python and JavaScript for more flexible use of the platform.

Considering the small amount of training data itself, the model is relatively simple. In this paper, a one-dimensional convolutional neural network shown in Figure 1 is used for model training. One-dimensional convolutional neural network (1D CNN) is a type of neural network based on convolutional operations, commonly used in the processing and analysis of sequence data. 1D CNN consists of various types of neural network layers, such as convolutional layers, pooling layers, and fully connected layers. In the field of audio classification, 1D CNN has been widely applied.



Input layer (3,960 features)

Reshape layer (40 columns)

1D conv / pool layer (8 neurons, 3 kernel size, 1 layer)

Dropout (rate 0.25)

1D conv / pool layer (16 neurons, 3 kernel size, 1 layer)

Dropout (rate 0.25)

Flatten layer

Add an extra layer

Output layer (4 classes)

Figure 1. The architecture of 1D CNN.

This article examines two distinct audio processing methods to identify the superior model for recognizing and classifying various audio types. The first method, Mel Frequency Cepstral Coefficients (MFCC) [9, 10], is based on Mel inversion coefficients. Initially, this technique segments the lengthy audio signal into smaller pieces and removes any edge effects. It then employs Fourier transform to generate a spectrogram, which is subsequently transformed into Meier cepstral coefficients. This conversion adjusts the frequency axis from a linear to a non-linear scale. Finally, the resulting coefficients are normalized to obtain the MFCC characteristics. Alternatively, in the MFE audio processing method, the spectrogram passes through a series of Meier filters, and the energy of the signal is captured. This energy value is then normalized to generate the MFE characteristics.

2.3. Implementation details

This article used the most common parameters. The whole training cycles is 100 and the learning rate is 0.005. This article uses both MFCC and MFE approaches to process audios. Both of them are some commonly used features extraction methods for speech signal processing. MFCC typically uses a discrete cosine transform (DCT) to convert the speech signal from the time domain to the frequency domain and compress the frequency domain data into a small, representative number of MFCC coefficients. These MFCC coefficients are typically used as input features in speech recognition systems. Unlike MFCC, MFE only considers the energy distribution of the audio signal without considering the information of frequency.

3. Result and discussion

Table 1 summarizes the results of multiple epochs of training conducted on an audio scene that employed two distinct audio processing methods, namely MFE and MFCC. A comparative analysis of the performance related to these two models revealed that using MFE processing method overall outperformed using MFCC processing method in terms of model recognition accuracy and training loss. Among them, when the number of predicted classes of classification method is the same, in the three-classification model, the accuracy of the training set of the MFE model is as high as 88.20%, while the accuracy of the training set of the MFCC model is only 84.90%; in the four-classification model, the accuracy of the training set of the MFE model is as high as 82.30%, while the accuracy of the training set of the MFCC model is as low as 77.60%.

Table 1. Performance of different CNNs.

		training performance		testing performance
		accuracy	loss	accuracy
MFE	3-classification	88.20%	0.39	80.61%
	4-classification	82.30%	0.65	60.49%
MFCC	3-classification	84.90%	0.44	74.04%
	4-classification	77.60%	0.64	53.66%

Next, the study conducted a comparison between the performance of a three-classification model and a four-classification model under the given scenario. Among the MFE and MFCC models, the triple classification model exhibited superior accuracy and lower loss. The losses were 0.39 and 0.44 when applying the triple classification method in both models, but the losses were substantially higher in the quadruple classification method, up to 0.65 and 0.64. The greater variation was in the performance of both models in the test set. The accuracy in the test set of the three-classification model is as high as 80.61% and 74.04%, which has some application value. In contrast, the recognition accuracy in the four-classification model was only 60.49% and 53.66%, which performed poorly.

According to most experimental laws, the recognition accuracy of the four-classification model should be higher than that of the three-classification model. However, in comic performances, it is often difficult to distinguish the actor's singing voice from the audience's cheering voice, and it is also difficult to distinguish the applause from the cheering voice. This is due to the high frequency of the performer's singing voice in this experimental application scenario, which is not clearly distinguished from the equally high frequency of the cheering voice, and the audience's applause is often mixed with the cheering voice, and the two often appear at the same time, leading to the lack of obvious training features. Therefore, in the comic scene, the triple classification MFE model is most appropriate.

4. Conclusion

In conclude, this study proposed different audio recognition to train crosstalk recognition models in low equality of data set using short clips from classical crosstalk including the sound of performers' sing, speaking and audiences' laughter, claps via deep learning to evaluate the quality of crosstalk by calculate the percentage of positive feedback. This article applied the MFCC and MFE audio processing method and 3-classification and 4-classification model to obtain the better performance of reaction audio recognition. The best model in this paper is 1D CNN with MFE method and 3-classification model, whose testing accuracy is up to 74.04%. In the future, this study will expand the number of data sets, trying to collect more different styles and times of crosstalk. This study will explore better models that can distinguish between singing and laughing sounds in order to classify high frequency audio signals exactly. There will be more CNN model applied in this research.

References

[1] Chen J 2014 Review of the Current Research Status of Modern Cross talk Art Journal of South China University of Technology: Social Science Edition 16 (5): 119-124

[2] Yang L Zhang Z 2021 Research on audio scene classification using improved convolutional neural networks Modern Electronic Technology 44 (3): 91-94

[3] Allamy S and Koerich A L 2021 1D CNN Architectures for Music Genre Classification 2021 IEEE Symposium Series on Computational Intelligence (SSCI) Orlando FL USA pp. 01-07

[4] Wang F Song G 2021 1D-TICapsNet: An audio signal processing algorithm for bolt early looseness detection Structural Health Monitoring 20(5):2828-2839

[5] Mohammad A U et al 2022 Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN Journal of Information and Telecommunication 6:1 27-42

[6] Aloysius N Geetha M 2017 A review on deep convolutional neural networks 2017 international conference on communication and signal processing (ICCSP) IEEE 0588-0592

[7] Rawat W Wang Z 2017 Deep convolutional neural networks for image classification: A comprehensive review Neural computation 29(9): 2352-2449

[8] Graves A Graves A 2012 Long short-term memory Supervised sequence labelling with recurrent neural networks 2012: 37-45

[9] Logan B 2000 Mel frequency cepstral coefficients for music modeling 270(1): 11

[10] Hasan M R Jamil M Rahman M 2004 Speaker identification using mel frequency cepstral coefficients variations 1(4): 565-568