Extraction of Broadband Vibration Spectrum Based on Audio-Visual Fusion

*CHENG Yao, YU Ruoyan, PENG Cong**

College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P. R. China

(Received 2 April 2022; revised 13 October 2022; accepted 21 June 2023)

Abstract: Vibration spectrum extraction is essential for fault diagnosis of rotating machinery. Environmental diversification and the presence of noise limit the performance of traditional single-modal vibration extraction methods. Since visual and audio signals have different sampling frequencies, noise and environmental constraints, audio-visual fusion can effectively solve the problem caused by single modality. Based on this, this paper proposes a wideband spectrum extraction method based on an audio-visual fusion deep convolutional neural network, which fully fuses the effective information of different modalities to complement each other. The proposed model uses a dual-stream encoder to extract features from different modalities, and a deep residual fusion module extracts high-level fusion features and feeds them to the decoder. The experimental results show that the performance of this model is superior to the latest vibration extraction methods, and the proposed model outperforms other state-of-the-art models such as RegNet, MFCNN, and L2L, which improves the accuracy of vibration spectrum extraction by 15% in noisy environment.

Key words: vibration spectrum extraction; audio-visual fusion; convolutional neural network; fault diagnosis; deep learning

CLC number: TP391.4 **Document code:** A **Article ID:** 1005-1120(2023)03-0323-13

0 Introduction

As one of the key components in rotating machinery, bearings are crucial for engines of aircraft and spacecraft. However, they are extremely susceptible to breakdowns and only very few of them reach their service life[1-2]. Bearing failure can cause major damage to the engines and in severe cases even devastating losses to the whole aircraft and spacecraft[3]. Vibration signature analysis has become one of the most used methods to identify possible rolling bearings failures. By calculating the bearing vibration frequency spectrum, the health status of the rotating machinery can be predicted and reflected[4].

Current vibration measurement methods are mainly divided into contact measurement and non-contact measurement. The contact measurement technology has been quite mature with high accuracy and reliability. However, the use of contact sensors to measure light structures inevitably causes mass loading effects by adding their weight to the original structure. In addition, in the scene of measurement for large structures, a large number of sensors need to be installed[5-6]. Moreover, the contact sensors can only perform a single-point measurement, invalid for full-field vibration extraction.

For non-contact measurement, vision-based vibration measurement has the advantages of low cost and is suitable for full-field measurement[7-9]. However, vision-based vibration measurement is easily affected by the environment, but not accurate under conditions such as limited illumination, complex target geometry, and constantly environments changing. More importantly, due to the influence of hardware technology and cost, it is difficult for digital

*Corresponding author, E-mail address: pengcong@nuaa.edu.cn.

How to cite this article: CHENG Yao, YU Ruoyan, PENG Cong. Extraction of broadband vibration spectrum based on audio-visual fusion[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2023, 40(3): 323-335.

http://dx.doi.org/10.16356/j.1005-1120.2023.03.008

cameras to measure vibrations in the middle and high-frequency bands[10]. As the camera sampling frequency increases, the hardware cost also increases rapidly.

Audio signals with a high sample rate are easily acquired. There exists a close correlation between acoustic and vibration signals. Many kinds of research are dedicated to recovering sound from vibration[11-13]. Zhao et al.[12] proposed a two-stream network to extract motion information from videos. A sound separation network was used to recover audio waveforms from the motion information and mixed audio. Mim et al.[13] proposed an optical flow method to extract vibration information from videos. The sound signal was recovered from that vibration signal. In addition, the fault diagnosis method of rotating machinery based on acoustics was also extensively studied. Yang et al.[14] proposed the stacked sparse encoders to detect the roller fault from the audio. Shan et al.[15] constructed an abnormal audio database. And they applied the wavelet packet algorithm and deep learning method to the diagnosis of partial discharge. Peng et al.[16] proposed the audio wavelet packet decomposition and convolutional neural network (CNN) for the fault diagnosis of rollers. Lu et al.[17] explored the potentials of the acoustic and vibration signals fusion and proposed a fault diagnosis method based on acoustic vibration signal fusion. Also, researchers explored acoustic signal processing methods for data enhancement of vibration signals[18-19]. Capturing high-frequency sound signals is convenient and the audio signal contains the vibration information of high frequency. However, all of these methods only indirectly exploit the relationship between audio and vibration. Due to the existence of a large amount of noise, it is difficult to accurately recover the vibration signal directly from the audio signal.

In recent years, multimodal fusion technology based on neural networks was widely used in many fields[20-23]. The fusion of different modalities can solve the difficulties of a single modality. Audio-visual fusion technology was developed for motion measurement, spectrum prediction, fault diagnosis, and other fields. Yoshida et al.[24] collected tennis swing videos from different athletes as an audio-visual motion dataset. The audio-visual fusion network was proposed to predict the acceleration information of the athlete's swing. Kist et al.[25] designed a high-speed video endoscope (HSV) system, which quantified vocal cord vibration information by analyzing the video and sound of the throat to assist doctors in the analysis of throat diseases. Hou et al.[26] proposed a rule-embedded network to fuse the audio-visual inputs for better detection of the target voice. The core role of the rule in the model is to coordinate the relation between the bi-modal information and use visual representations as a mask to filter out the information of non-target sound. Sari et al.[27] introduced a multi-view model that uses shared classifiers to map audio and video into the same spatial fusion for high-performance speaker recognition tasks. Lu et al.[28] calculated the rotor speed and angle values from continuous video frames and the envelope spectrum curve of the rotor fault audio. They resampled the envelope spectrum according to the angle values to obtain the fault feature envelope spectrum. And the obtained spectrum was compared with the theoretical fault sequence spectrum for fault diagnosis. In addition, audio-visual fusion is widely used in tasks such as speech enhancement and prediction[29-30]. Tao et al.[29] proposed a novel end-to-end, multitask learning (MTL), audiovisual ASR (AV-ASR) system. A key novelty of the approach is the use of MTL, where the primary task is AV-ASR, and the secondary task is audiovisual voice activity detection (AV-VAD). Tzirakis et al.[30] developed an emotion recognition system that utilizes the raw text, audio and visual information in an end-to-end manner. The system utilize novel attention-based methods to fuse the modality-specific features.

However, most audio-visual fusion methods select the compressed spectrum with a narrow bandwidth as the frequency domain features[31]. The high-frequency information of the compressed spectrum is limited and incomplete. The occurrence of early faults often exists in the form of high frequency in the task of fault diagnosis. Low-frequency information is often concealed by noise and is difficult to

find[32].

This paper proposes an audio-visual fusion network for vibration spectrum extraction in bearing fault diagnosis. The proposed network can extract broadband spectrum to obtain the high-frequency vibration information of objects. Moreover, the proposed network extract features in video and audio stream separately through the dual-stream encoder network. Subsequently, an audio-visual fusion module containing fully connected layers and residual convolution blocks is used to fuse the features. Finally, the broadband frequency spectrum is reconstructed through the decoder. The audio-visual datasets GRID, UCF101, and DCASE are used to jointly construct audio-visual broadband spectrum datasets for training and testing. To validate the performance of the method in the application of fault diagnosis, the pre-trained model is used to conduct test experiments on the bearing fault audio-visual dataset.

The main contributions of this paper can be summarized as follows:

(1) We proposed audio-visual residual fusion model(AVRF), which is a multimodal fusion algorithm more suitable for extracting broadband vibration signals from videos.

(2) We adopt the spatial-temporal convolution (STC) in the encoder, which filters out the pixels with low relevance to audio input and enhances the correlation of audio and visual features. The proposed structure also leads to the improvement of model performance without the region of interest processing.

The rest of the paper is organized as follows. Section 1 introduces the framework and structure of the proposed model. Section 2 illustrates the employed datasets and audio-visual signal preprocessing method. Section 3 describes the experimental results, and the conclusion part is shown in Section 4.

1 Model Architecture

In this section, a broadband vibration spectrum extraction method is proposed. AVRF is used to extract broadband vibration signals from videos. Fig.1 shows the algorithm flow based on audio-visual fusion broadband spectrum extraction. AVRF uses a two-stage fusion strategy to recover vibration information from videos and acoustic signals with low signal-to-noise ratios. Features in convolutional layers can only affect each other in a limited range. Therefore, using convolutional layer fusion only would result in insufficient fusion. The first-stage fusion uses fully connected layers to fuse signals of different modalities. Shallow networks are incapable of dealing with complex and varied broadband outputs. The residual structure avoids the problem of network degradation while increasing the network layers. To output the broadband vibration spectrum accurately, the residual structure is introduced in the second-stage fusion. Using the residual structure in the fusion module has better performance. Because the effective information is mainly extracted in the feature fusion stage.

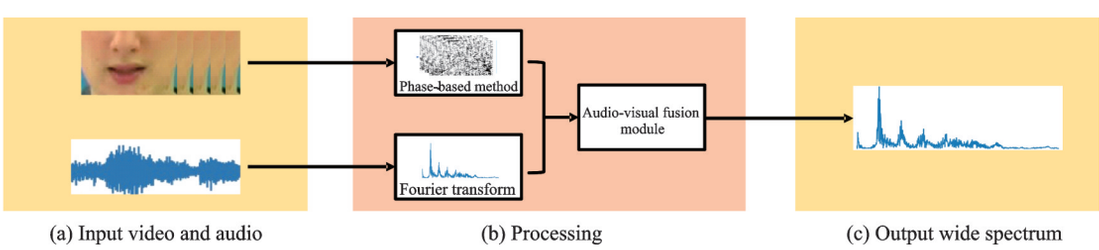

Fig.1 Audio-visual fusion broadband spectrum extraction algorithm flow

The proposed model involves an audio encoder, video encoder, audio-visual fusion module, and decoder. The network structure of the four parts is shown in Fig.2.

1.1 Audio encoder

Concerning many previous models involving audio encoders, the audio encoder used in this paper is also composed of some convolutional layers, and

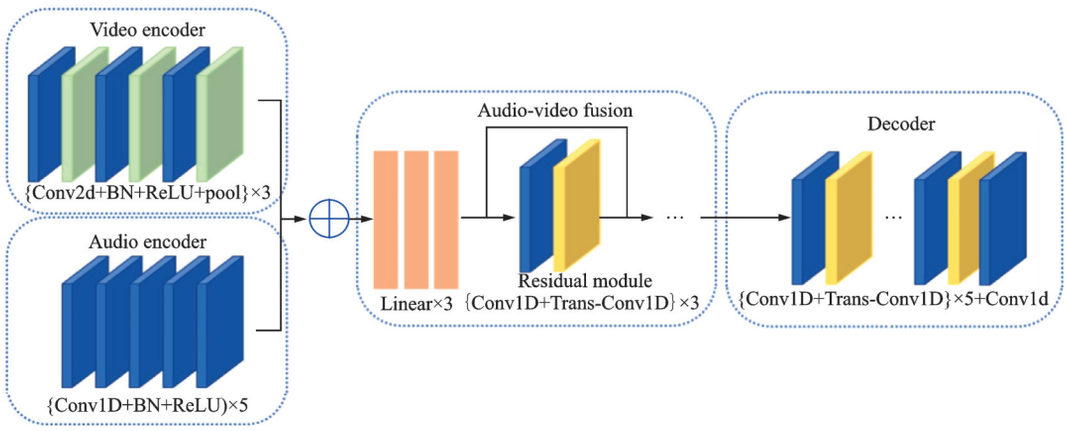

Fig.2 Illustration of the proposed audio-visual fusion model

the audio spectrum is used as the audio encoder input.

Specifically, the audio encoder is composed of five one-dimensional convolutional layers. Each convolutional layer is followed by batch-normalization and the ReLU function for non-linearity.

1.2 Video encoder

The video encoder uses a phase-based method to extract visual temporal information in consecutive video frames as input. Considering the input three-dimensional feature X, where X is extracted from input video by the phase based method. Its dimension is image width, image height, and frame rate. They represent two spatial dimensions and a temporal dimension. The phase changes in the temporal dimension almost involve the most frequency information in the visual features. Therefore, the spatial-temporal convolution (STC) is performed by a video encoder on the input temporal dimension. The number of convolution kernels is gradually reduced to compress the spatial information. As a result, only the features of useful pixels in the spatial dimension are selected for fusion with audio features, and the features of pixels with low relevance to audio information are filtered out.

The video encoder has six layers, which are alternately composed of a two-dimensional convolutional layer and a two-dimensional pooling layer. Each convolutional layer is followed by batch-normalization and the ReLU function is used for non-linearity.

1.3 Audio-visual fusion module

At the audio-visual fusion stage, a consolidated dimension is usually selected to integrate multimodal features, in which the visual features and audio features are concatenated. In this paper, the temporal dimension is chosen to merge visual and audio features. Although the sampling rate of audio information is much higher than that of video information, the importance of different modal information should be similar. In the fusion process, different modal information should be given the same initial weight. Therefore, by designing the network structure of the video encoder and audio encoder, the output of the visual and audio features from the encoder are flattened at the same feature dimension, and then a serial operation is performed on this temporal dimension.

The audio-visual fusion module includes two parts. Liking the soft attention mechanism, the attentional mask is often obtained through the fully connected layer. The first part is composed of three fully connected layers. The function of this part allows the visual and audio features to fully integrate. Compared with the convolutional layer, the fully connected layer can sufficiently give different weights to the audio-visual features, highlight the important information and filter out unimportant ones and noise.

The second part is composed of three residual fusion modules. The parameters of the residual fusion module are described in Table 1. It is well known that the deeper the layers are, the more abstract the features and semantic information are ex-

tracted. In this paper, the input of the model also contains the broadband spectrum with a large amount of information.

Table 1 Detailed architecture of residual fusion module

Module	Number of filter	Filter size	Stride
Conv1	128	5	1
Trans-Conv1	64	3	1
Conv2	128	5	1
Trans-Conv2	128	3	1
Conv3	256	5	1
Trans-Conv3	256	3	1

It is difficult for the shallow network to extract abstract high-level features. But simply increasing the convolutional layers may cause gradient disappearance, gradient explosion, and network degradation. The residual structure can effectively alleviate the above problems. Using the residual module for fusion here can not only extract higher-level fusion information but also make network prediction more adaptable to spectrum prediction tasks of different bandwidths.

1.4 Decoder

The decoder consists of five groups of convolutional layers and transposed convolutional layers. Each convolutional layer is followed by batch-normalization and the ReLU function for non-linearity. Parameters of the decoder are the mirror setting of the audio encoder. The detailed parameter is shown in Table 2.

Table 2 Detailed architecture of the decoder

Module	Number of filter	Filter size	Stride
Conv1	256	5	1
Trans-Conv1	256	3	2
Conv2	64	5	1
Trans-Conv2	64	3	1
Conv3	16	3	1
Trans-Conv3	16	3	2
Conv4	8	3	1
Trans-Conv4	8	4	2
Conv5	8	3	1
Trans-Conv5	8	4	2
Conv6	1	3	1

2 Dataset and Preprocessing

2.1 Dataset

The dataset in this paper is a combination of two public audio-video datasets. The first is the GRID[33] audio-visual speech dataset. The audio-visual data includes speech fragments of 34 people (18 men and 16 women). There are 1 000 video clips for each person, which contain audio and videos of the speaker's face. The second dataset is UCF101[34], which is an audio-visual action one. It contains 13 320 video clips from 101 action categories. This dataset provides the greatest diversity in the action of objects and great changes in camera position, object appearance, etc. This paper chooses to use five instrument categories of the video. For the videos in these two datasets, the region of interest (ROI) regions are cropped out and these cropped regions contain the main motion information in the videos. Fig.3 shows the cropped representative frames of these videos.

In addition, the audio scene classification dataset DCASE[35] is used as the noise. This dataset contains scene categories such as bus stops, busy streets, parks, markets, offices, restaurants, subway stations, etc. Each category has 10 audios which are mainly used to add noise to audio input. The clean audio from GRID, UF101, and the noise audio from DCASE are resampled to the same frequency. The mixed audio is obtained by adding noise to clean audio. Mixed audio has a signal-to-noise ratio of -5 — 0 dB by controlling the energy level of the noise. Fig.4 shows the audio waveform of clean audio, noise audio and mixed audio. Take 1 s of audio data as a sample, and the dataset has a total of about 12 000 samples. Eighty percent of them are used for training and the rest are used for testing. The noise added to the training set and the test set is from different categories.

In addition, a bearing audio-visual dataset is used to examine the effectiveness of AVRF for bearing vibration spectrum extraction[28]. The test bearing is driven by a brushless direct current motor car-

Fig.3 Cropped video segments of 64 pixel×64 pixel in the dataset

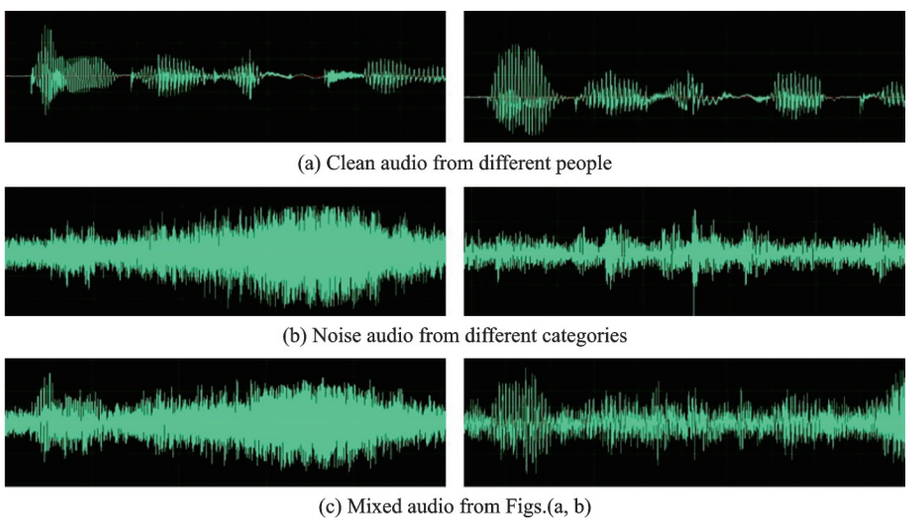

Fig.4 Audio waveform visualization in the dataset

rying a mechanical load through a coupling. The visual and audio signals of the mechanical bearing are collected by an iPhone 5S synchronously. The video frame rate is captured at 240 frame/s and the audio signal is sampled at 44 100 Hz. The system collects the audio-visual data of the bearing under an inner raceway fault to verify the validity of the proposed model in extracting the information of the fault vibration spectrum. The pre-trained model is directly used for testing in this dataset. The video and audio data are processed the same as the GRID and UF101 dataset to fit the pre-trained model.

2.2 Data preprocessing

There are two processing methods for video data. The first is to crop the area of interest (such as the mouth of speakers, the part of the hand-played instrument, etc.) to a size of 64 pixel×64 pixel. The second method does not crop the frames but resizes the video frames to 64 pixel×64 pixel. It is a common preprocessing method to crop the video frame and obtain the ROI as the visual input of the network in the speech signal processing task of audio-visual fusion. The raw audio wave is mixed with the original noise signal, and the signal-to-noise ratio after mixing is controlled to be −5—0 dB.

The video data uses a phase-based method[36] to extract temporal information. The frame rate of the video is uniformly resampled to 25 frame/s. A total of 50 frames are selected as the video stream input, which is the size of 64×64×50. Audio features are extracted by FFT as spectral information which is fed to the audio encoder. It can be divided into three different width spectrums 0—100 Hz, 0—500 Hz, and 0—1 000 Hz. By combining two classes of videos input, four experiments with different types of input and output are designed.

3 Experiment Results and Discussion

3.1 Setup

To evaluate the performance of the proposed audio-visual fusion broadband spectrums extraction model, this paper uses an audio-only model and several audio-visual fusion models for comparison.

These models all use the audio spectrums as the output, and the tasks of these networks include speech enhancement, video sound generation, etc. We use the following models to compare both quantitatively and qualitatively. TCNN[37] is a temporal convolutional network based on audio-only. It is a speech enhancement algorithm which extracts clean spectrum from audio input with noise. It is mainly used for comparison with audio-visual fusion models. L2L[38] is an audio-visual fusion network for speech separation tasks. This audio-visual model extracts a single speech spectrum from a mixture of sounds such as other speakers and background noise. MFCNN[39] is an audio-visual fusion network for speech enhancement tasks. It has an encoder-decoder structure with the multi-scale fusion strategy. This model can remove the noise components of the input spectrum by audio-visual fusion. RegNet[40] is an audio-visual joint training network for the prediction of sound from video. This model is designed to predict the frequency spectrum of an object in the input video.

The experimental results used the following evaluation metrics to evaluate model performance. Mean squared error (MSE) is used for measuring the distance between spectrums. Logarithmic spectrum distance (LSD) and log-likelihood ratio (LLR) show the overall similarity of the spectrums. The signal-to-noise ratio (SNR) and the segmented signal-to-noise ratio (SegSNR) weigh the energy level of the effective signal and the noise signal.

3.2 Comparison with audio-only model

To determine the significance of audio-video fusion, the temporal convolutional neural network (TCNN) is contrasted. It is an audio-only model and the result can be found in Table 3. TCNN has a similar structure to the proposed model in this paper, consisting of an encoder-decoder structure and a time-domain convolution module (TCM) with a residual structure.

Two experiments are conducted with the output bandwidths of 500 Hz and 1 000 Hz. In each experiment, the audio-only model shares the same audio input as the AVRF. In the 500 Hz experiment,

Table 3 Performance of the audio-only model TCNN and the proposed AVRF

Bandwidth/Hz	Model	MSE	LSD	SNR	LLR	SegSNR
500	TCNN	0.016	0.956	1.203	1.491	0.235
	AVRF	0.007	0.947	2.933	1.571	1.427
1 000	TCNN	0.013	1.014	0.814	1.623	0.819
	AVRF	0.009	0.987	2.143	1.532	2.229

the AVRF has a better performance than TCNN in most evaluation metrics. Moreover, in the case of 1 000 Hz, the proposed model even gets a better improvement in metrics compared with the audio-only model. This proves that the multi-modal fusion process can select more effective information from the multiple modal features. The audio-video fusion method can extract a more accurate spectrum than the single-modal method.

The qualitative comparison results in Fig.5 also show the extraction of frequency values by TCNN and AVRF. The red dot points out the peak value predicted at the wrong position. The spectrum of TCNN has four frequency peaks at 107, 214, 238, and 321 Hz. Three of them are

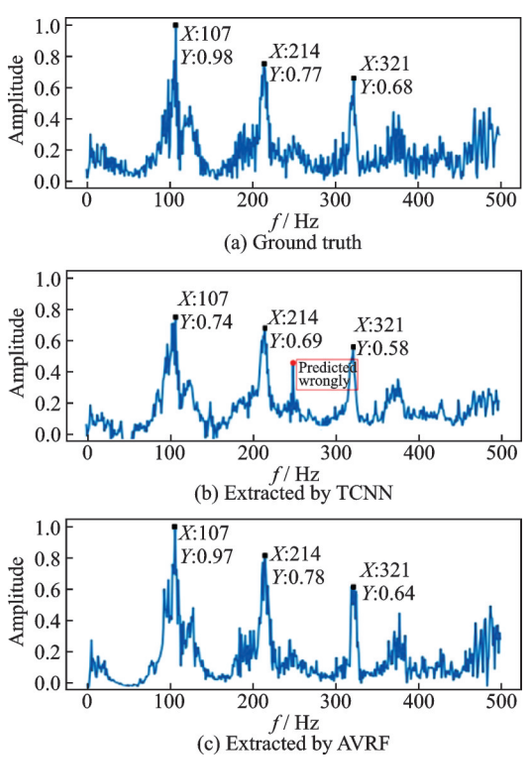

Fig.5 Comparison results of the extracted spectrum

the same as the peaks in the ground truth. However, the peak of 238 Hz is completely predicted wrongly. The frequency peaks in the spectrum of AVRF and the ground truth are the same at 107, 214, and 321 Hz. In addition, the frequency amplitude of AVRF is closer to the ground truth than that of TCNN. Therefore, the qualitative comparison proves that the AVRF model can extract a more accurate broadband spectrum than the audio-only model.

3.3 Comparison with audio-video fusion models

It is commonly used short-time Fourier transform (STFT) or Mel cepstral transform for audio feature preprocessing in tasks such as speech recognition and speaker recognition. These methods compress the spectrum over frequency bands to represent deep semantic features. The frequency resolution of them is usually selected as 80 or 160. Therefore, many audio-visual fusion networks are limited in outputting spectrum with a high-frequency resolution.

A subjective comparison test is conducted to examine the effectiveness of the proposed model. Three experiments perform with the output bandwidth of 100, 500, and 1 000 Hz. In the 100 Hz experiment (Table 4), the spectrum extracted by MFCNN outperforms the others in evaluation metrics. Other comparative models have also achieved similar performance, and the proposed model has no superiority in narrow-band spectrum extraction.

Table 4 Performance of audio-video fusion models with 100 Hz bandwidth output

Model	MSE	LSD	SNR	LLR	SegSNR
L2L	0.002 2	0.398 0	16.061	0.869	15.772
RegNet	0.005 4	0.578 9	11.639	0.732	13.254
MFCNN	0.003 1	0.466 7	16.150	0.634	18.492
AVRF	0.007 2	0.636 4	9.661	1.518	10.742

In the 500 and 1 000 Hz experiment (Tables 5, 6), the AVRF has achieved a clear improvement with a 25% reduction of MSE scores, 10% reduction of LSD scores, 18% increase of SNR and SegSNR than the state of the art models. As new structures have been added to the AVRF, including STC in video encoder and the deep residual fusion module. It can be observed that models with deeper network layers (such as L2L) have a relatively better performance than MFCNN and RegNet. Hence the improvement of AVRF is primarily for two reasons, the STC in video encoder instead of convolution only in the spatial dimension and the deep residual fusion module, which has a stronger semantic representation ability. In addition, the proposed model also has a relatively better inference complexity, which is 209 giga floating point operations (GFLOPs) compared with RegNet with 386 GFLOPs, L2L with 238 GFLOPs and MFCNN with 83 GFLOPs.

Table 5 Performance of audio-video fusion models with 500 Hz bandwidth output

Model	MSE	LSD	SNR	LLR	SegSNR
L2L	0.015	0.864	1.751	1.465	1.324
RegNet	0.022	1.067	0.482	1.554	0.587
MFCNN	0.021	1.047	0.611	1.652	0.655
AVRF	0.007	0.947	2.933	1.571	1.427

Table 6 Performance of audio-video fusion models with 1 000 Hz bandwidth output

Model	MSE	LSD	SNR	LLR	SegSNR
L2L	0.012	0.952	1.896	1.611	1.352
RegNet	0.021	1.172	0.482	2.203	0.232
MFCNN	0.022	1.085	0.611	1.865	1.127
AVRF	0.009	0.987	2.143	1.532	2.229

A qualitative comparison between models is shown in Fig.6. The colored dashed lines mark the main frequency peaks, red arrows indicate the missing peak prediction, green arrows point out the wrong frequency amplitude, and orange arrows point out the peak value predicted at the wrong position. The 500 Hz bandwidth output of the different models was visualized and compared as a spectrum. Frequency peaks have greater importance in the vibration spectrum for containing more vibration information. In bearing fault diagnosis, the frequency peak of bearing vibration can be used as effective information for judging the mechanical health status. Therefore, it is significant that the vibration extraction model can accurately output all the frequency

peaks.

The output from MFCNN and RegNet is almost impossible to accurately predict major frequency peaks. Similar to the comparison results of the evaluation metrics, L2L accurately extracts the first three main frequency peaks. However, L2L also extracts additional non-existent frequency peaks at the wrong location. The AVRF model obtains results consistent with the ground truth, in which the frequency and amplitude of the peak are extracted accurately. The arrows of different colors indicate the frequency peak redundancy, the frequency peak missing, and the frequency peak mismatch in the output spectrum of the comparison model.

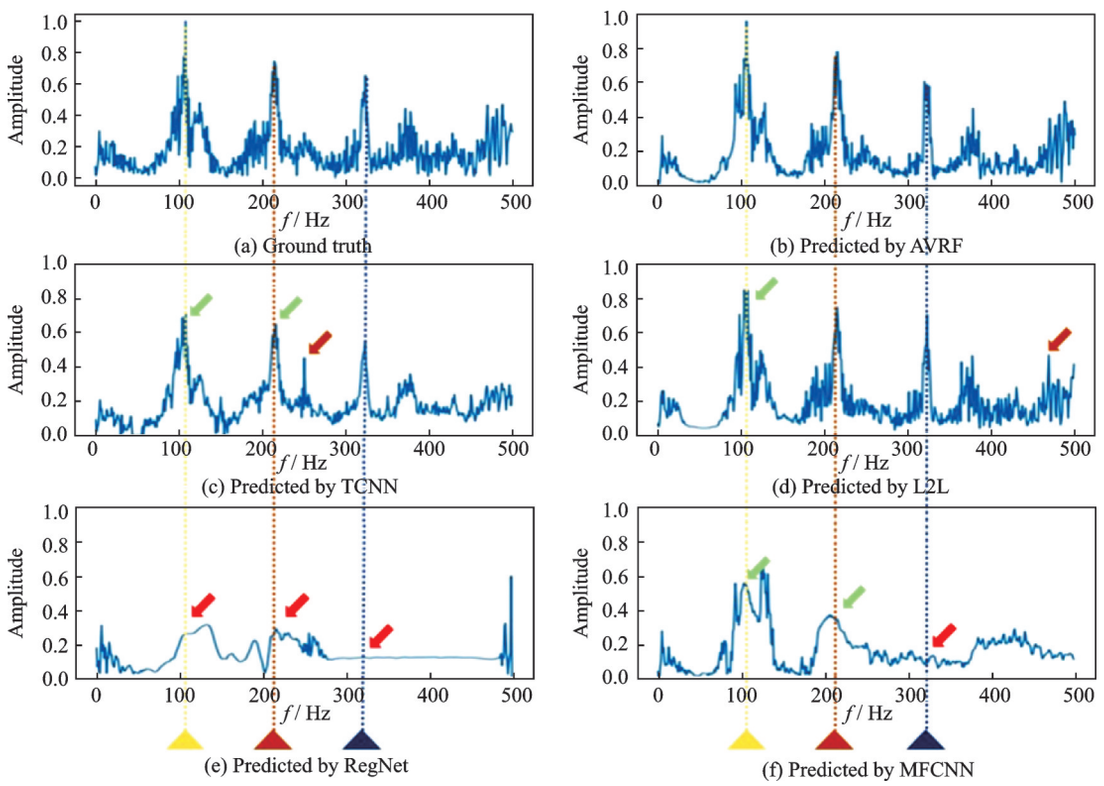

Fig. 6 Comparison results of the extracted spectrum

3.4 Comparison between cropped and resized visual input

The major previous audio-visual fusion models show a limited feature extraction capability of the visual processing module. In these models, the visual encoder is usually a pre-trained model with limited depth compared to a complete visual feature extraction network. Therefore, cropping the video frame to obtain ROI can reduce irrelevant visual input during the fusion process. However, selecting the ROI usually requires additional detection algorithms, which leads to additional computational cost and affects the real-time performance of the method, especially in mechanical fault diagnosis tasks.

The AVRF filters out irrelevant information in visual modality features through our audio-visual fusion module based on the correlation between audio-visual signals. In speech enhancement, emotion recognition, and other tasks, the RGB features of video frames are often used as visual modal input. In the vibration extraction task, audio-visual modal features have a stronger correlation, which both represent the motion information. The phase-based method is used to extract vibration information in video frames, and each pixel in the visual input represents a motion sequence, which is also one of the reasons why the audio-visual fusion method can effectively filter out irrelevant information.

To demonstrate the capability of AVRF for processing uncropped visual input, we test all models on cropped and resized visual input comparatively. The experimental results of cropping the ROI or not

are compared and the results are shown in Table 7. It can be found that the compared models have a decline in the performance of the LSD and LLR metrics when the ROI is not selected. The model in this paper has almost no decline in these two indicators, indicating that the proposed model achieves a better fusion of features from different modalities. The proposed audio-visual fusion module can utilize the correlation between audio features and visual features for better selecting the visual information to a certain extent.

Table 7 Performance of all the models with 500 Hz spectrum and no cropping visual input

Model	Preprocessing	MSE	LSD	LLR
TCNN	Cropping	0.015	0.956	1.491
	No cropping	0.016	0.903	2.013
L2L	Cropping	0.015	0.864	1.465
	No cropping	0.149	0.965	1.847
RegNet	Cropping	0.019	1.067	1.554
	No cropping	0.022	1.014	1.869
MFCNN	Cropping	0.021	1.047	1.652
	No cropping	0.018	1.160	1.651
AVRF	Cropping	0.007	0.947	1.571
	No cropping	0.007	0.938	1.588

3.5 Test on bearing dataset

The vibration of rotating machinery has the characteristics of high frequency, small amplitude, and periodicity. The natural frequency and high-frequency harmonics of bearing vibration can reflect the state of rotating machinery. The AVRF model has a better performance in extracting high-frequency vibration information. The visual modal features can accurately extract small-amplitude vibration information, and audio modal features contain high-frequency and periodic vibration information. Therefore, the AVRF model is suitable for bearing vibration extraction.

To examine the effectiveness of the proposed model for bearing vibration spectrum extraction, a test on a bearing audio-visual dataset is presented. The results are shown in Table 8. The vibration signal has a more effective energy level than that of the speech or musical instrument signals. Therefore, the signal-to-noise ratio metrics are better in the rotor dataset experiment. The proposed model has achieved similar performance for the rest metrics on two different types of the test dataset. The proposed model demonstrates a robust performance on broadband vibration spectrum extraction.

Table 8 Performance of the AVRF on different datasets

Dataset	MSE	LSD	SNR	LLR	SegSNR
GRID and UF101	0.007	0.947	4.140	1.571	4.427
Bearing dataset	0.011	0.730	6.060	1.316	6.302

Fig.7 shows the spectrum visualization of the proposed model and the ground truth. The AVRF model obtains similar results with the ground truth, and the frequency of the peak is extracted accurately. The comparison results demonstrate that AVRF can be adapted to the task of extracting the wideband vibration spectrum of the mechanical bearing. In addition, it can be seen that the bearing vibration spectrum has more frequency peaks and less background noise energy, which is significantly different from the speech and instrument. This indicates that AVRF has stronger adaptability to different types of vibration extraction.

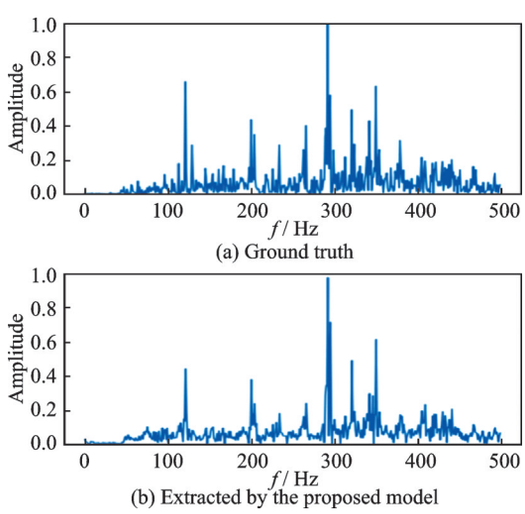

Fig.7 Comparison results of the extracted spectrum on bearing audio-visual dataset

4 Conclusions

This paper proposes a broadband vibration spectrum extraction model based on audio-visual fusion, which can be applied to extract the vibration spectrum of rotating machinery. The audio and visu-

al features are extracted by the audio encoder and the video encoder respectively, and the outputs of encoders are merged by the audio-visual fusion module composed of fully-connected layers and multiple residual modules to obtain the audio-visual hybrid features. Finally, the fused features are extracted to a broadband spectrum through the decoder.

Compared with the state-of-the-art audio-visual fusion spectrum extraction models, the model in this paper has a better performance in multiple broadband prediction experiments on the joint data of GRID and UCF101. The proposed model reduces the spectrum extraction error by 25% and improves the accuracy of vibration spectrum extraction by 15% in noisy environment. By comparing the generated spectrograms, the broadband spectrum extracted by AVRF is closer to the ground truth, the position and amplitude of frequency peak in the spectrum are almost accurately predicted. In addition, the pre-trained model on the GRID and UCF101 joint dataset are tested on the bearing audio-visual dataset, and it also achieved a good performance on multiple evaluation indicators. For future studies, the attention mechanism can be used in the audio-visual fusion stage to achieve a more balanced fusion representation, by filtering out unimportant information and some noise.

References

[1] YU F, WEI D, ZUO H, et al. Time sequence change-point model of electrostatic state parameters of aircraft engine[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2018, 35(1): 130-138.

[2] CUI J, TIAN Y, CUI X, et al. An effective fault diagnosis method for aero engines based on GSA-SAE[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2020, 37(5): 755-757.

[3] DING B, WU J, SUN C, et al. Sparsity-assisted intelligent condition monitoring method for aero-engine main shaft bearing[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2020, 37(4): 508-516.

[4] ISHAM M F, LEONG M S, MENG H L, et al. Variational mode decomposition for rotating machinery condition monitoring using vibration signals[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2018, 35(1): 38-50.

[5] FAN W, QIAO P. Vibration-based damage identification methods: A review and comparative study[J]. Structural Health Monitoring, 2011, 9(3): 83-111.

[6] STANBRIDGE A B, EWINS D J. Modal testing using a scanning laser doppler vibrometer[J]. Mechanical Systems & Signal Processing, 1999, 13(2): 255-270.

[7] PENG C, ZHU M, WANG Y, et al. Phase-based video measurement for active vibration suppression performance of the magnetically suspended rotor system[J]. IEEE Transactions on Industrial Electronics, 2021,68(2): 1497-1505.

[8] PENG C, ZENG C, WANG Y. Phase-based noncontact vibration measurement of high-speed magnetically suspended rotor[J]. IEEE Transactions on Instrumentation and Measurement, 2019, 69(7): 4807-4817.

[9] PENG C, ZENG C, WANG Y. Camera-based micro-vibration measurement for lightweight structure using an improved phase-based motion extraction[J]. IEEE Sensors Journal, 2019, 20(5): 2590-2599.

[10] BUSCA G, CIGADA A, MAZZOLENI P, et al. Vibration monitoring of multiple bridge points by means of a unique vision-based measuring system[J]. Experimental Mechanics, 2014, 54(2): 255-271.

[11] DAVIS A, RUBINSTEIN M, WADHWA N, et al. The visual microphone: Passive recovery of sound from video[J]. ACM Transactions on Graphics, 2014, 33(4): 1-10.

[12] ZHAO H, GAN C, MA W, et al. The sound of motions[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). [S.l.]: IEEE, 2019: 1735-1744.

[13] MIM K Z, MIAH A A, AHMAD M. Extraction of sound signal from tiny vibrations in motion magnified video using optical flow[C]//Proceedings of the International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). [S.l.]:IEEE, 2019.

[14] YANG M, ZHOU W, SONG T. Audio-based fault diagnosis for belt conveyor rollers[J]. Neurocomputing, 2020, 397: 447-456.

[15] SHAN H, JING L, HUAN Q, et al. Acoustic diagnosis of partial discharges in transformers[C]//Proceedings of the 2020 IEEE Conference on Electrical Insulation and Dielectric Phenomena (CEIDP). [S.l.]: IEEE, 2020.

[16] PENG C, LI Z P, YANG M, et al. An audio-based

intelligent fault diagnosis method for belt conveyor rollers in sand carrier[J]. Control Engineering Practice, 2020, 105: 104650.

[17] LU S, ZHENG P, LIU Y, et al. Sound-aided vibration weak signal enhancement for bearing fault detection by using adaptive stochastic resonance[J]. Journal of Sound and Vibration, 2019, 449: 18-29.

[18] MDESA M. Audio deformation based data augmentation for convolution neural network in vibration analysis[J]. IOP Conference Series: Materials Science and Engineering, 2019, 551: 012066.

[19] ESA M F, MUSTAFFA N H, OMAR H, et al. Learning convolution neural network with shift pitching based data augmentation for vibration analysis[J]. IOP Conference Series: Materials Science and Engineering, 2020, 864: 012086.

[20] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37: 98-125.

[21] ZHANG J H, YIN Z, CHEN P, et al. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review[J]. Information Fusion, 2020, 59: 103-126.

[22] CHANDRASEKARAN G, NGUYEN T N, JUDE H D. Multimodal sentimental analysis for social media applications: A comprehensive review[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2021. DOI:10.1002/widm.1415.

[23] THOMBRE S, ZHAO Z, RAMM-SCHMIDT H, et al. Sensors and AI techniques for situational awareness in autonomous ships: A review[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(1): 64-83.

[24] YOSHIDA K, HORIUCHI Y, ICHIYAMA T, et al. Estimation of racket grip vibration from tennis video by neural network[J]. Haptic Interaction: Perception, Devices and Algorithms, 2019, 535: 33-45.

[25] KIST A M, GOMEZ P, DUBROVSKIY D, et al. A deep learning enhanced novel software tool for laryngeal dynamics analysis [J]. Journal of Speech Language and Hearing Research, 2021, 64(6): 1889-1903.

[26] HOU Y B, DENG Y, ZHU B, et al. Rule-embedded network for audio-visual voice activity detection in live musical video streams[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S. l.] : IEEE, 2021: 4165-4169.

[27] SARI L, SINGH K, ZHOU J T, et al. A multi-view approach to audio-visual speaker verification[C]// Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).[S.l.]:IEEE, 2021: 6194-6198.

[28] LU S, JIE G, HE Q, et al. A novel contactless angular resampling method for motor bearing fault diagnosis under variable speed[J]. IEEE Transactions on Instrumentation & Measurement, 2016, 65(11): 2538-2550.

[29] TAO F, BUSSO C. End-to-end audiovisual speech recognition system with multitask learning[J]. IEEE Transactions on Multimedia, 2020, 23: 1-11.

[30] TZIRAKIS P, CHEN J X, ZAFEIRIOUS S, et al. End-to-end multimodal affect recognition in real-world environments[J]. Information Fusion, 2021, 68: 46-53.

[31] ABDELAZIZ A H. Improving acoustic modeling using audio-visual speech[C]//Proceedings of IEEE International Conference on Multimedia & Expo (ICME2017). [S.l.]: IEEE, 2017: 1081-1086.

[32] NGUYEN P, KANG M, KIM J M, et al. Robust condition monitoring of rolling element bearings using de-noising and envelope analysis with signal decomposition techniques[J]. Expert Systems with Applications, 2015, 42(22): 9024-9032.

[33] COOKE M, BARKER J, CUNNINGHAM S, et al. An audio-visual corpus for speech perception and automatic speech recognition[J]. The Journal of the Acoustical Society of America, 2006, 120(5): 2421-2424.

[34] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. Computer Science, 2012. DOI: 10.48550/arXiv.1212.0402.

[35] MESAROS A, DIMENT A, ELIZALDE B, et al. Sound event detection in the dcase 2017 challenge[J]. ACM Transactions on Audio, Speech, and Language Processing,2019, 27(6): 992-1006.

[36] WADHWA N, RUBINSTEIN M, DURAND F, et al. Phase-based video motion processing[J]. ACM Transactions on Graphics, 2013,32(4): 80-89.

[37] PANDEY A, WANG D L. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain[C]//Proceedings of ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 6875-6879.

[38] EPHRAT A, MOSSERI I, LANG O, et al. Looking

to listen at the cocktail party: A speaker-independent audio-visual model for speech separation[J]. ACM Transactions on Graphics, 2018, 37(4): 112.1-112.11.
[39] XU X M, WANG Y, XU D X, et al. MFFCN: Attention multi-layer feature fusion convolution network for audio-visual speech enhancement[EB/OL].(2022-09-27). https://doi.org/10.48550/arXiv.2101.06268.
[40] CHEN P, ZHANG Y, TAN M, et al. Generating visually aligned sound from videos[J]. IEEE Transactions on Image Processing, 2020, 29: 8292-8302.

Acknowledgements This work was supported by the National Science Foundation of China(No.62122038) and the Natural Science Foundation of Jiangsu Province (No. BK20211565).

Authors Mr. **CHENG Yao** was born in Heilongjiang, China, in 1998. In 2016, he received the B.S. degree in College of Automation Engineering from Nanjing University of Aeronautics and Astronautics. He is studying for the M.S. degree at Nanjing University of Aeronautics and Astronautics. His current research interests include audio-visual fusion, vibration measurement and deep learning.

Prof. **PENG Cong** received the B.S. degree in instrument science and technology from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in instrument science and technology from Beihang University, Beijing, China, in 2016. She is currently a professor with College of Automation Engineering, Nanjing University of Aeronautics and Astronautics. Her research interests include active vibration control, vibration measurement and artificial intelligence.

Author contributions Mr. **CHENG Yao** designed the model, conducted the analysis, interpreted the results, and wrote the original manuscript. Ms. **YU Ruoyan** contributed to the model validation and helped to perform the analysis with constructive discussions. Prof. **PENG Cong** supervised this study, polished English writing of the manuscript and contributed to the discussion and revision of the study. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing inrests.

(Production Editor: SUN Jing)

基于视听融合的宽频带振动频谱提取

程 遥，于若颜，彭 聪

(南京航空航天大学自动化学院,南京 211106,中国)

摘要:提取振动频谱对于旋转机械的故障诊断至关重要。环境和噪声的多样化限制了传统单模态振动提取方法的性能。由于视听信号具有不同的采样频率、噪声和环境限制,视听融合算法可以有效解决单一模态存在的问题。基于此,文中提出了一种基于视听融合深度卷积神经网络的宽带频谱提取方法,该方法充分融合了不同模态的有效信息。该模型基于双流编码器从不同的模态中提取特征,使用深度残差融合模块提取高级融合特征并输出给解码器。实验结果表明,该模型的表现优于最新的振动提取方法,如RegNet, MFCNN及L2L等,噪声环境下的振动频谱提取准确率提高15%。

关键词:振动频谱提取;视听融合;卷积神经网络;故障诊断;深度学习