Detecting Aging Substation Transformers by Audio Signal with Deep neural network

Wei Ye[1], Jiasai Sun[1], Min Xu[1], Xuemeng Yang[2], Hongliang Li[2*], and Yong Liu[2]

[1] State Grid Zhejiang Electric Power Company's information communication company, Zhejiang, China
[2] Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, China

Abstract. In order to monitor the aging of transformers and ensure the operational safety in substations, a practical detection system for indoor substation transformers based on the analysis of audio signal is designed, which use computer technology instead of manpower to efficiently monitor the transformers working states in real-time. Our work consists of a small and low cost AI-STBOX and an intelligent AI Cloud Platform. AI-STBOX is installed directionally in each transformer room for continuously collecting, compressing and uploading the transformers audio data. The AI Cloud Platform receives audio data from AI-STBOX, analyses and organizes the data to low-dimensional speech features with STFT and Mel cepstrum analysis. Input the features into a powerful deep neural network, the system can quickly distinguish the working states of each substation transformer before is has serious faults. It can locate aging transformers, command the maintenance platform to quickly release the repair task, thus avoid unforeseeable outages and minimize planned downtimes. The approach has achieved excellent results in the substation aging transformers detection scene.

Keywords: Substation transformer · speech feature · deep neural network.

1 Introduction

The normal operation of the 10kV distribution line is of great significance to the power system, which is also an important guarantee to meet the electricity demand of urban and rural residents in China. However, it is undeniable that at present, there are still many problems in operating the 10kV distribution lines, which seriously affects the operation and management of distribution lines and is not conductive to the development of electric power industry. The 10kV distribution network is mostly radially arranged and varies in length. There can be many branch units in one line including switching station, high-voltage branch boxes, transformers, low-voltage cabinets, low-voltage branch boxes, etc., reaching dozens or even hundreds of units, and the connections are even more numerous. The 10kV distribution lines have very complicated paths, and the quality of the equipment is also uneven, so they have different aging rates. Coupled with the influence of external factors, it is difficult to find and deal with the fault quickly in the event.

Transformers are the cornerstone of power generation, transmission, and distribution systems, while transformer aging may brings many adverse effects such as transformer winding burnout, transformer insulation breakdown. Once a fault occurs, it may results in power outage of a specific district or even more serious consequence. Through investigation, we find existing substation monitoring methods rely on manpower. When inspecting the power equipment, electricity workers often judge the states of transformers through experience. The judgment process is actually closely related to the frequency of the sound heard by human. The ability of fault detection is obtained by long-term experience accumulation, that is, some special audio modes correspond to abnormal states. However, it is unrealistic to require patrol personnel to inspect substations frequently as the distance between substations is long and some substations are in remote areas. Such intermittent inspections cannot detect substation abnormalities in real-time and effectively, thus failing to deal with equipment faults in time, resulting in a large risk.

According to the inspection methods of the electricity workers, the aging equipments produce different sounds with normal equipments, which can be an alert of the serious malfunction. The feasibility of audio analysis and the need of real-time inspection inspired us to use a deep learning network to monitor the aging of equipments.

In this work, we adopt an audio classification technology through deep neural network to distinguish transformers states. It is one of the key technologies on how to deal with, analyse and utilize massive audio information. Before classification, any audio signal needs to be extracted its features, which can be classified into time-domain features and spectrum features. Time-domain analysis uses the waveform of the audio signal itself for analysis. Spectral analysis uses the spectral representation of the audio signal for analysis. Common time domain features include Volume Distribution[16], Pitch Contour[16], zero-crossing rate[7], short-time energy [6], short-time autocorrelation function, short-time average amplitude difference, etc. Common frequency domain features include: short-time Fourier transform (STFT)[2], wavelet transform[4], ST[22], etc. There are many feature extraction techniques, including Linear Predictive Analysis (LPC)[8] Linear Prediction Cepstral Coefficient (LPCC)[1], Perceptual Linear Prediction Coefficient (PLP)[10], Mel Frequency Cepstral Coefficient (MFCC)[17], Power Spectrum Analysis (FFT), Relative Spectral Transform of Log Domain Coefficients (RASTA)[11], First Derivative (DELTA).

The development of audio classification technology has a long history, and a series of methods have emerged. In 1977, Sawhney and Maes from MIT Media Lab [21] recorded a dataset from a set of classes including people, voices, subway, traffic and others. Then employed recurrent neural networks and a k-nearest neighbour criterion to model the mapping between the extracted features and categories. Stan Z. Li proposed a new pattern classification method based on a so-called nearest feature line (NFL) in 2000[15]. Select at least two different samples in each audio category, and make a cepstrum feature of any two samples into a straight line in the feature space, called a feature line. The query input is compared with each feature line separately, and the category of the closest feature line is selected as the query result. Then in 2003, the author proposed a content-based audio classification and retrieval method based on support vector machine (SVM)[9]. Given a feature set consisting of cepstrum features, learn the best class boundaries between classes from the training data by using SVM and finally match by using the distance from the boundary. Recent years, more researchers [3] [13] [14] [12] used deep network models to classify sounds in the environment. Karol J. Piczak[18] evaluated the potential of convolutional neural networks in classifying short audio clips of environmental sounds. Gerard Roma [19] described a new set of descriptors based on Recurrence Quantification Analysis (RQA), and proposed a framework for environmental sound recognition based on blind segmentation and feature aggregation. These approaches make it possible to classify sound with neural networks for substation transformers.

In this work, a complete online detecting system for indoor substation transformers is introduced, which collects transformers audio data with a non-contact way, extracts its features and judges the current working state of every transformer efficiently by a deep network. It not only saves a lot of unnecessary human resources, but also achieves the purpose of monitoring substation status in real-time.

2 System Structure

Inspired by traditional manual detection methods, we analyse the feasibility of monitoring the transformers by sounds. Besides, sound detection is a non-contact detection method, which can ensure its safety when using in substations. We install AI-STBOX in each transformer room, and the audio data is continuously collected for a certain length of time, compressed and uploaded to the AI Cloud Platform. The AI Cloud Platform analyses and organizes the audio data, extracts the audio features it contains, inputs which into the deep network that we have trained, and finally obtains the results, reflecting the real-time status of each transformer room(aging or not). The whole structure is shown in Figure 1.

2.1 Introduction of AI-STBOX

AI-STBOX is an audio data acquisition and transmission equipment designed in this work which is shown in Figure 2. It is installed in transformer rooms for collecting, compressing and transmitting audio data. As we adopt a supercardioid pointing dynamic microphone in AI-STBOX, it directionally captures the sound in front of the microphone and ignores sounds from other directions.Therefore, the noise in the scene is avoided to some extent. AI-STBOX combines a powerful 1.2Ghz multi-core processor to make calculations smoother and more responsive.

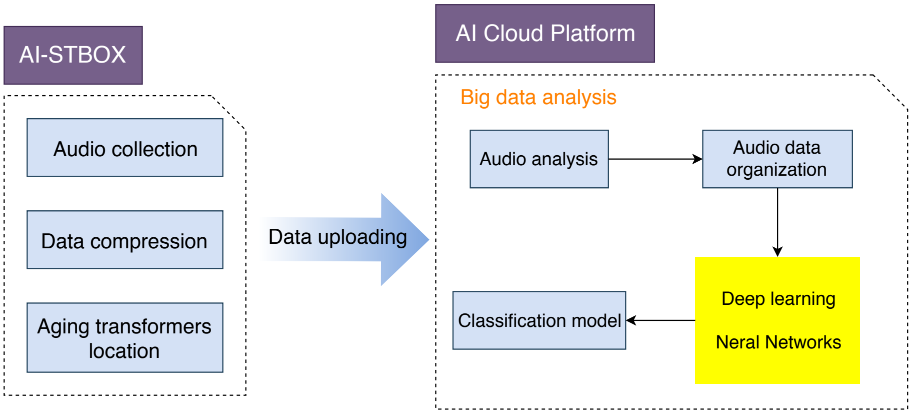


Fig. 1. Introduction of technical solutions. AI-STBOX can upload compressed audio data to AI Cloud Platform, and AI Cloud Platform analyses the data to monitor if the transformer is aging.

The integrated high-performance DTU module is equipped with 2G/3G/4G network for more reliable data transmission. It is small and low-cost, meet the needs of practical applications in the factory. Moreover, it is industrialized design, dustproof, waterproof and shockproof, and can monitor transformer equipment 24 hours a day in real-time.

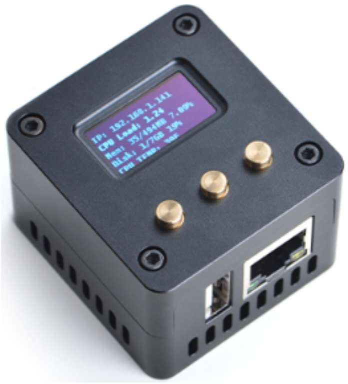


Fig. 2. AI-STBOX, an audio data acquisition and transmission equipment.

The device is based on an open architecture system design with portability, interoperability, and tailorability. Portability indicates that various computer applications can be ported between various computer systems with open structural features, no matter they are of the same type or model. Interoperability means that the nodes on the computer network can interoperate and share resources, no matter the nodes are of the same type or model. The tailorability shows that the application system running on the low-end machine of the system should be able to run on the high-end machine, and the application system running on the high-end machine can also be run on the low-end machine after being cut. This allows people to view the running status of the low-end machine on the high-end machine. If any AI-STBOX is damaged, we can check and replace it in time to make the whole system more stable.

We install and power up the AI-STBOX in the transformer room and place it towards the transformers. The AI-STBOX continuously collects audio data of the substation in a fixed long time. For example, the audio is stored once a minute, compressed and transmitted to the detection platform.

2.2 AI Cloud Platform

The AI Cloud Platform is used to receive and analyse the raw audio data of each transformer room, and finally identify whether it reflects an aging equipment. The platform mainly includes audio feature extraction, deep network classification, post-processing of detection results.

Audio Feature Extraction After AI-STBOX transmits the original audio data to the AI Cloud Platform, it is first input to the audio feature extraction module to extract the distinguishing features. The original sound signal is a one-dimensional time domain signal, and it is difficult to visually see the frequency variation pattern. Although the frequency distribution of the signal can be transformed by Fourier transform into the frequency domain , the time domain information is lost so that the change of the frequency distribution with time cannot be seen. In order to solve this problem, many time-frequency analysis methods have emerged. Short-time Fourier transform (STFT) is the most commonly used time-frequency analysis method. It represents the signal characteristics at a certain moment by calculating the signal in time window. The speech signal is time-varying, but at a short time interval, it can be assumed that the signal has hardly changed or changed very little, and the Fourier transform is performed on each time window of the signal to obtain a spectrogram.

To be specific, for the acquired long-time sound data, we seperate the data to several short-time level frames. For each frame, use sliding window to obtain the very short-time interval. For each very short-time interval, performing short-time Fourier transform to get a high-dimensional spectral feature. Short-time Fourier transform is shown as

$$X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi k/N}, 0 \leq \mathrm{k} \leq \mathrm{N} \tag{1}$$

where N is the length at Fourier transform, $x(n)$ is the signal to be transformed, X_a is the Fourier transformation result. Figure 3 shows a sample of spectrogram generated by Short-time Fourier transform.

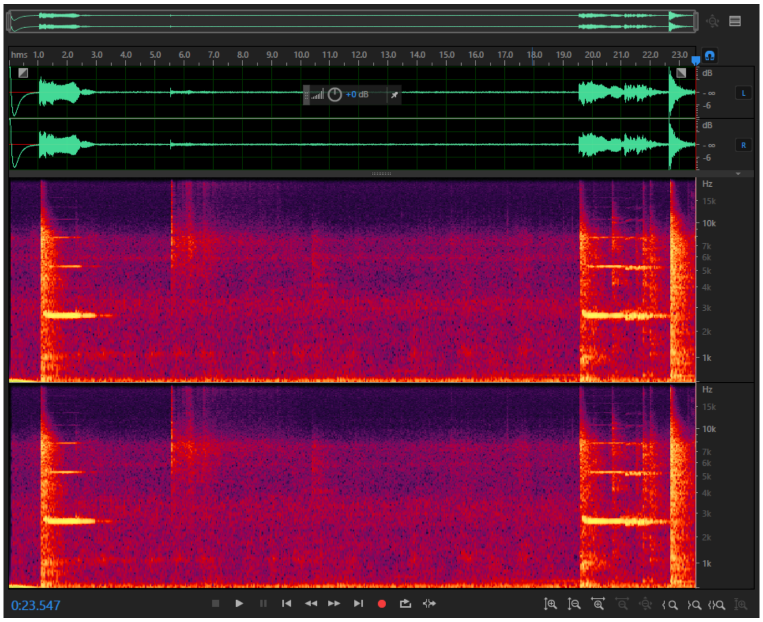


Fig. 3. Audio waveform and spectrogram of an abnormal electric device

As humans' perception of Hertz is not a linear perception, to make it more similar to the human auditory system, the obtained spectrum is usually passed through a Mel filter banks to obtain the Mel spectrum. Several band-pass filters $H_m(k)$ $(0 \leq m \leq M)$, are set, where M is the number of filters. Each filter has a triangular filtering characteristic, and its center frequency is $f(m)$. In the range of Mel frequency, these filters are of equal bandwidth. The transfer function of each band-pass filter is

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{2}$$

$f(m)$ can be defined as

$$f(m) = \left(\frac{K}{f_s}\right) F_{mel}^{-1}\left(F_{mel}\left(f_l\right) + m\frac{F_{mel}\left(f_h\right) - F_{mel}\left(f_l\right)}{M+1}\right), \tag{3}$$

where f_l is the lowest frequency of the filter frequency range, f_h is the highest frequency of the filter frequency range, f_s is the sampling frequency, F_{mel} function is as

$$F_{mel} = 1125 \ln(1 + f/700). \tag{4}$$

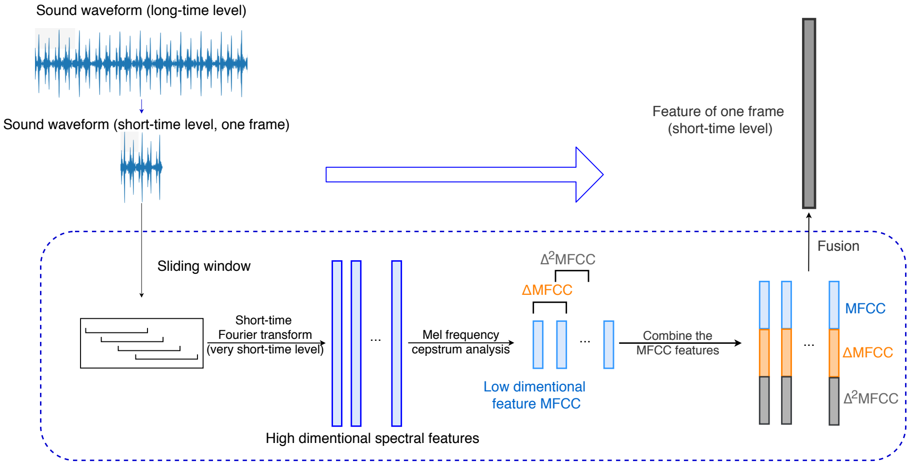


Fig. 4. The process of speech feature extraction. We extract the high-dimensional spectral features of each short-time audio frame and transform them to MFCC, ΔMFCC and Δ^2MFCC. Combine the MFCC features to obtain the final feature.

After the obtained spectrum is passed through the Mel filter bank, perform cepstrum analysis on the Mel spectrum. Specifically, take the logarithm of the powers that we obtained from M Mel filter banks

$$\mathrm{s(m)} = \ln\left(\sum_{k=0}^{N-1} \left|X_a(k)\right|^2 H_m(k)\right), 0 \leq \mathrm{m} \leq \mathrm{M}. \tag{5}$$

then perform the inverse Fourier transform, which is generally realized by discrete cosine transform(DTC). Here we use a type of DCT as

$$\mathrm{C(n)} = \sum_{m=0}^{N-1} \mathrm{s}(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1, 2, \ldots, L. \tag{6}$$

In the field of sound processing, Mel frequency cepstrum is a linear transformation of the logarithmic energy spectrum of a non-linear Mel scale based on sound frequency. The difference between cepstrum and Mel frequency cepstrum is that the frequency division of the Mel frequency cepstrum is equally spaced on the Mel scale, which is more similar to the human auditory system than the linearly spaced frequency band used in the normal logarithmic cepstrum. Such a non-linear expression makes it possible to have a better representation of the sound signal in multiple fields.

Use the 2nd to 14th coefficients after DCT as the Mel-frequency cepstral coefficients (MFCC)[20].The MFCC is the coefficient that constitutes the Mel frequency cepstrum which is widely used for speech recognition. It is proposed by Davis and Mermelstein in the 1980s[5] and has continued to be one of the most advanced technologies. Taking the MFCC composed of 13 coefficients, the difference of the MFCC as ΔMFCC and the difference of the ΔMFCC as Δ^2MFCC, and combine these three parameters to obtain the speech feature representation corresponding to the audio frame. Short-time speech features are composed of a plurality of very short-time MFCC speech feature expressions. Figure 4 shows the process of speech feature extraction.

The MFCC, ΔMFCC and Δ^2MFCC are both of 13-dimensional, and the combination feature corresponding to the very short time is of 39-dimensional. Either compared with the original audio data or the high-dimensional spectral features, the amount of speech feature data corresponding to the very short time is greatly reduced, and thus the calculation amount of the subsequent deep learning classification module is reduced. Then, through a combination of a plurality of very short-time speech features, short-time speech features are obtained.

Deep Neural Network Classification In this section, the deep neural network is used to learn the mapping relationship between sound features and classification results. As we have shown the possibility to judge the transforms' status by sounds, the deep neural network is a replacement of human to carry out this work. The open-source deep network framework Pytorch is used to build a deep network of 3 fully-connected layers, dropout and batch normalization are also added to improve the generalization ability of the network and accelerate convergence. The structure of the network is shown in Figure 5.

The pre-acquired and labeled short-time audio features obtained from the substation are used for training, and the label of each short-time audio feature is consistent with the long-time audio to which it belongs. The normal substation sound data is marked as 1, while abnormal sound data is marked as 0. After the input data passes through the audio feature extraction model, the audio features are sent to the deep network in batches, so that the depth model can gradually learn the mapping relationship between the input audio features and the output evaluation. The last layer of the network is connected to the Sigmoid function to ensure the classification result is in range $(0, 1)$. Sigmoid function is as

$$S(t) = \frac{1}{1 + e^{-t}}. \tag{7}$$

The output represents the normal probability corresponding to the input feature. The output value close to 1 indicates that the short-time audio is normal, the working state of the substation tends to be normal. The output value close to 0 means that the short-time audio tends to be abnormal.

Post-processing of Test Results In order to obtain a more robust result, the output of the deep neural network is encapsulated in a higher level. The deep network output indicates the detection result of relatively short-time audio data. The post-processing part is to combine multiple short-time audio detection results output by the deep networks to obtain the abnormal detection result

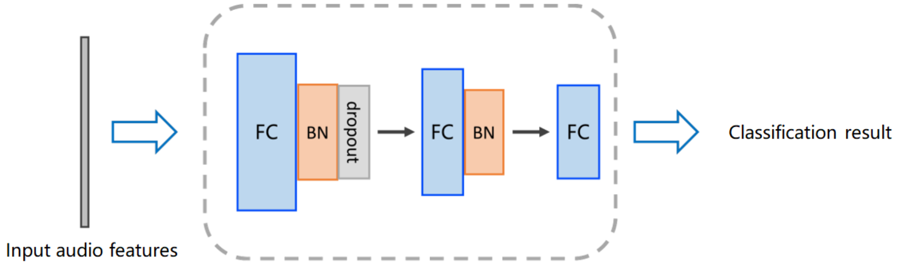


Fig. 5. Deep neural network structure. FC represents the fully connected layer, and BN represents the batch normalization.

corresponding to a long-time audio, thereby improving the reliability of the detection method.

Specifically, for each long-time audio we detect the status of multiple short-time audio separately, and when the number of abnormal estimated results is greater than a set threshold, the detection result is determined to be abnormal, and the threshold may make adjustment according to a specific scenario. If the test results for several consecutive long-time intervals indicate that the current transformer is abnormal, it may need to be reported to the relevant maintenance department and check.

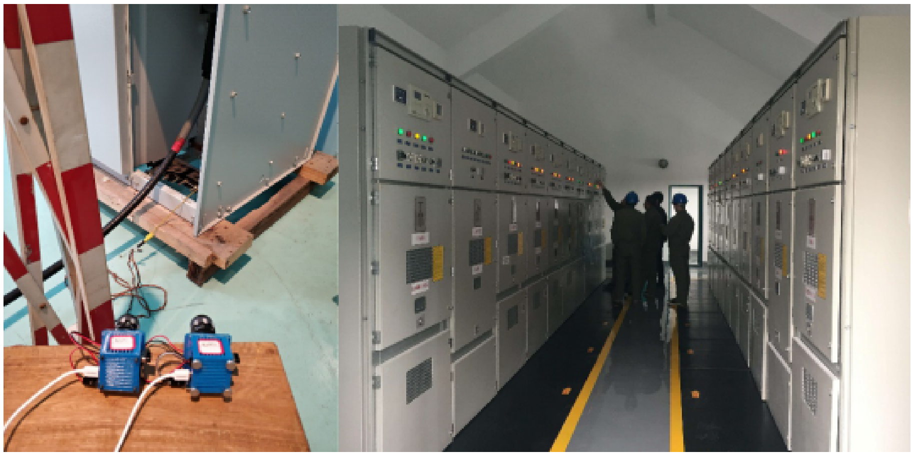


Fig. 6. Scene of collecting audio data

3 Experiments

3.1 Experimental Details

AI-STBOXs are directionally installed to collect data with a sampling rate at $f_s = 8$kHz. They provide the on-site audio, which are stored and transmitted to the AI Cloud Platform once a minute. In order to achieve a smoother input, in the process of sampling the original audio data, set the sliding window length of audio frames as 64ms, moves the sliding window every 16ms. Each sliding window is only 1/4 of the frame length, ensuring consistent feature coverage. The feature input to the network for training is obtained by splicing 50 MFCCs, whcih means the final input to the network is a feature corresponding to a duration of (16ms*50)-16ms+64ms=848ms. Finally, the long audio of 1 minute can be split into about 60s/0.848s=73 short-time samples. Set $f_l = 300$Hz, $f_h = 4$kHz, $M = 26$ and $K = 512$ in the Mel filter bank. The deep neural network is built on the open-source framework Pytorch 0.3, and the structure of the network is based on 3 fully connected layers. In order to ensure an efficient training process, the speech features corresponding to all audio frames in the training data are extracted in advance, and the features are directly input to the network.

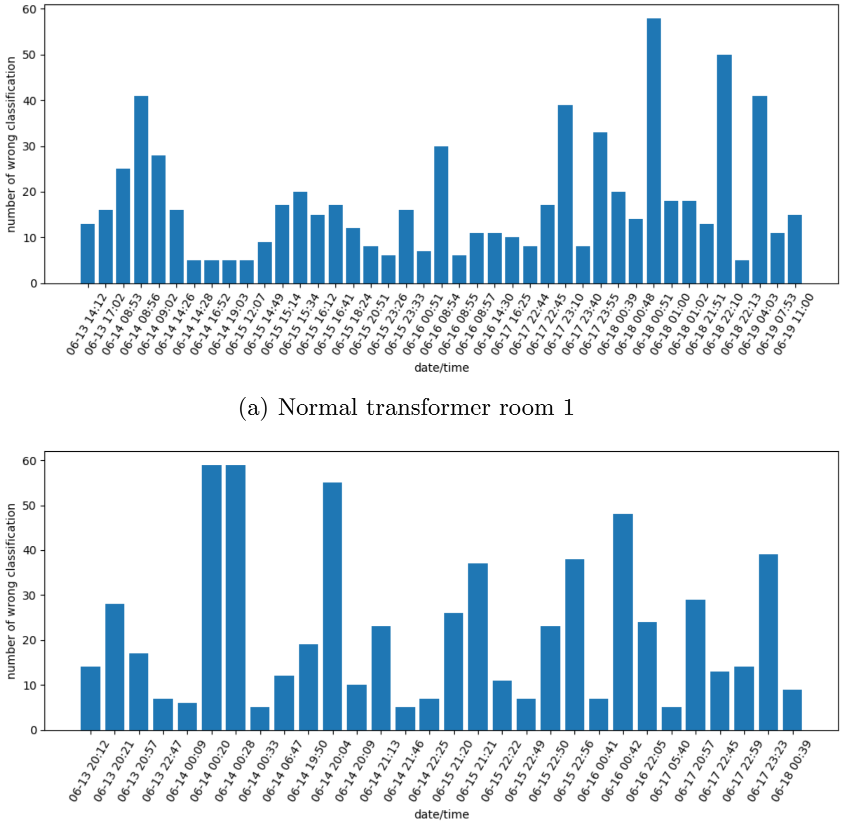
During the test, 73 short-time frames in one minute are sent to the deep learning network for detection separately. When the number of abnormal states is greater than 5, the state of the current minute can be considered abnormal. When the status of the transformer for 3 consecutive minutes occurs abnormality, it is considered that the maintenance personnel can be notified to check the transformer.

3.2 Dataset

The audio data for experiments are collected in multiple indoor substations for several days. One of the scene of collecting data is shown in Figure 6. Label the normal or abnormal state of the collected substation data, including the label for each long-time audio. The collected data from aging equipment is abnormal data which marked as 0. The collected data from newly-input substation is normal data which marked as 1. The audio and its corresponding tags are stored for later training of deep learning classification models. A total of about 770 hours of data was collected, and about 337w audio samples were generated, of which 80% were used as training data and the remaining 20% were test data. The wav file is an audio file developed by Microsoft Corporation that supports multiple compression algorithms. The audio files used in this experiment were saved in wav format.

3.3 Results

Take 20% of the data as a test set and the rest as training set. The ratio of positive and negative samples is about 7 to 3. The test set includes a total of 158.82 hours of data, which is divided into 674,250 short-time samples. The



(a) Normal transformer room 1

(b) Normal transformer room 2

Fig. 7. Time distribution of the evalued results in normal transformer room 1(a) and 2(b). The histogram shows the number of wrong classification for short-time samples in the corresponding minute.

overall test time is 9.13 minutes. In the 674,250 short-time samples, only 6052 samples are evalued wrongly. The overall accuracy is 99.10%, recall is 99.66%, and precision is 99.10%. The average test of one minute audio data only takes 1.2ms, the real-time and accuracy can meet the abnormal detection needs of the substation.

The estimation results of short-time samples in each minute are counted. When the results of 5 or more than 5 short-time samples are abnormal, the prediction of this minute is judged to be abnormal. We select the examples in which the actual state is normal but judged wrongly. The histogram for two transformer rooms are shown in Figure 7 (test results for two rooms respectively). It can be seen that the estimation result will not cause an error in several consecutive minutes. We set the threshold time to the maintenance notification is 3 minutes, which increase the fault tolerance and stability of our system, avoiding frequent overhauls and the waste of human resources.

The experimental results on the test set also show that within the 9238 minutes data used in the test, only 4 minutes of abnormal state is divided into normal state. It shows that our method is robust and can be applied in engineering and has practical significance if the samples are sufficient and the amount of data is large enough.

4 Conclusions

In this work, we design a substation transformers online detection system based on the analysis of audio. The system uses a non-contact way to collect transformers sounds and aims to effectively judge aging transformers by sounds through deep neural network, which achieves timely feedback. Through this method, aging equipments can be located, serious accidents can be predicted and avoided to some extent. This makes accident response time greatly shortened, and has significant economic benefits. The results demonstrate the feasibility of applying the proposed method in practical scene.

References

1. Ai, O.C., Hariharan, M., Yaacob, S., Chee, L.S.: Classification of speech dysfluencies with mfcc and lpcc features. Expert Systems with Applications **39**(2), 2157–2165 (2012)
2. Allen, J.B., Rabiner, L.R.: A unified approach to short-time fourier analysis and synthesis. Proceedings of the IEEE **65**(11), 1558–1564 (1977)
3. Barchiesi, D., [email protected], G., Stowell, D., Plumbley, M.: Acoustic scene classification: Classifying environments from the sounds they produce. Signal Processing Magazine, IEEE **32**, 16–34 (05 2015). https://doi.org/10.1109/MSP.2014.2326181
4. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE transactions on information theory **36**(5), 961–1005 (1990)
5. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences (1980)
6. Enqing, D., Guizhong, L., Yatong, Z., Yu, C.: Voice activity detection based on short-time energy and noise spectrum adaptation. In: 6th International Conference on Signal Processing, 2002. vol. 1, pp. 464–467. IEEE (2002)
7. Gouyon, F., Pachet, F., Delerue, O., et al.: On the use of zero-crossing rate for an application of classification of percussive sounds. In: Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy. p. 26 (2000)
8. Gray, R.M.: Linear Predictive Coding and the Internet Protocol. Now Publishers (2010)
9. Guo, G., Li, S.Z.: Content-based audio classification and retrieval by support vector machines. IEEE transactions on Neural Networks **14**(1), 209–215 (2003)
10. Hermansky, H.: Perceptual linear predictive (plp) analysis of speech. the Journal of the Acoustical Society of America **87**(4), 1738–1752 (1990)
11. Hermansky, H., Morgan, N.: Rasta processing of speech. IEEE transactions on speech and audio processing **2**(4), 578–589 (1994)

12. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine **29**(6), 82–97 (Nov 2012). https://doi.org/10.1109/MSP.2012.2205597
13. Kons, Z., Toledo-Ronen, O.: Audio event classification using deep neural networks. pp. 1482–1486 (01 2013)
14. Lee, H., Pham, P., Largman, Y., Ng, A.Y.: Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in neural information processing systems. pp. 1096–1104 (2009)
15. Li, S.Z.: Content-based audio classification and retrieval using the nearest feature line method. IEEE Transactions on Speech and Audio Processing **8**(5), 619–625 (2000)
16. Liu, Z., Wang, Y., Chen, T.: Audio feature extraction and analysis for scene segmentation and classification. Journal of VLSI signal processing systems for signal, image and video technology **20**(1-2), 61–79 (1998)
17. Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In: ISMIR. vol. 270, pp. 1–11 (2000)
18. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6 (Sep 2015). https://doi.org/10.1109/MLSP.2015.7324337
19. Roma, G., Herrera, P., Nogueira, W.: Environmental sound recognition using short-time feature aggregation. Journal of Intelligent Information Systems **51**(3), 457–475 (2018)
20. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech Communication **54**(4), 543–565 (2012)
21. Sawhney, N., Maes, P.: Situational awareness from environmental sounds. Project Rep. for Pattie Maes pp. 1–7 (1997)
22. Stockwell, R.G., Mansinha, L., Lowe, R.: Localization of the complex spectrum: the s transform. IEEE transactions on signal processing **44**(4), 998–1001 (1996)