An approach for monitoring sand mining based on sound feature

Qingmei Zhou¹ and Xiping He¹,²,³*

¹National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing 400067, China
²School of Artificial Intelligence, Chongqing Technology and Business University, Chongqing 400067, China
³Chongqing Engineering Laboratory for Detection, Control and Integrated System, Chongqing Technology and Business University, Chongqing, 400067, China
*Corresponding author’s e-mail: jsjhxp@ctbu.edu.cn

Abstract. In order to strengthen the management of sand excavation in river courses and prevent the occurrence of illegal sand mining activities, this paper proposes an approach for monitoring sand mining based on sound. Firstly, Mel Frequency Cepstral Coefficients (MFCCs) abstractor and Autoencoder are combined to extract features of every frame of a sound sequence, and then each frame of the sound sequence is classified by a specific classifier. Finally a voting strategy is used among the frames to determine the final category of the sound sequence. Experiments show that whether the classifier is SVM, KNN, or BP neural network, the result of combined features is better than the result of features extracted by the MFCC abstractor. Therefore, it is feasible to use the artificial intelligence method based on sound features extracted through MFCC abstractor and Autoencoder to monitor sand mining.

1. Introduction
River (lake) sand is a very good building material, and its market demand is quite huge, resulting in frequent illegal sand mining activities, so the effective control of sand mining activity has become an urgent problem to be solved by river management departments. Considering that sound contains a lot of useful information, and the process of collecting sound is simple and the cost of collecting is low, a method of monitoring sand mining based on sound is proposed in this paper.

In today's sound analysis, the most common acoustic features are MFCCs [1]. For example, MFCCs and two parallel feed forward neural networks are used to classify aircraft [2]; MFCCs and minimum mean-square error (MMSE) are used to evaluate speech features of the underlying noise-free speech signal for automatic speech recognition (ASR) [3]; MFCCs and extreme learning machine are used to classify excavation equipment [4]. These papers first use the MFCC abstractor to extract features of sound, and then perform further sound analysis. However, MFCCs are hand-crafted feature representation [5]. Studies have shown that there are always different types of noise mixed in the actual sound environment. Even if noise is limited within a narrow frequency band, the noise effect will still spread over all MFCCs [6]. So hand-crafted features may not always be able to adapt to complex application scenarios in practice [5].

Autoencoder can automatically capture the most important factor that represents the input data and find the main component that can represent the original information [7], and be robust to the noise,
especially non-stationary noise [8]. Therefore, this paper considers using MFCC abstractor to extract the initial features, and then uses Autoencoder as the second sound features learning model in order to enhance the performance of feature learning.

The rest of this paper is organized as follows. This paper firstly introduces the technology roadmap of monitoring sand mining based on sound in Section 2, of which 2.1 introduces the data preprocessing before extracting the sound features, 2.2 introduces how to extract the features of sound, and 2.3 describes the process of classification. Section 3 shows the experimental setup and results. Section 4 concludes the paper.

2. Technology roadmap
The technology roadmap of monitoring sand mining based on sound features includes three steps: pretreatment, feature extraction, and classification.

2.1. Pretreatment
In order to extract audio features smoothly, the audio signal must be preprocessed firstly. The process is as follows:

(1) Data preparation: First of all, the acquired audio files are converted into audio signal data. Then, multi-channel is converted to mono. All audio signals are resampled to 16KHZ. Every audio file is split into many more audio files, each with the length of 2s.

(2) Pre-emphasis: Pass the sound signal through a first-order high-pass filter:

\[ y[n] = x[n] - a \times x[n-1] \]  

(1)

Where \( a \) is between 0.9 and 1.0, and this paper takes 0.97. The purpose of pre-emphasis is to boost the energy in the high-frequency part so that the spectrum can be obtained with the same signal-to-noise ratio over the entire frequency band.

(3) Framing: In order to facilitate sound analysis, one sound sequence can be divided into consecutive small segments, that is, \( N \) sample points are grouped into a unit of observation called frame. In order to avoid excessive changes in adjacent frames, an overlapping area is left between adjacent frames. In this paper, each frame contains 128 sample points, with 64 sample points overlapping between adjacent frames, and the time length of each frame is 16ms.

2.2. Feature extraction

2.2.1. Using MFCC abstractor to initially extract features
Mel frequency cepstral coefficients, as the name suggests, feature extraction involves two key steps: transforming to Mel frequency, then cepstral analysis. The correspondence between the frequency(Hz) and the Mel scale is

\[ mel(f) = 1127 \log(1 + \frac{f}{700}) \]  

(2)

The flow chart for extracting MFCCs is shown in figure 1, and each step is explained as:

![Sound Signal -> Windowing -> FFT -> Mel filtering -> Logarithm -> DCT -> Cepstrum](image)

Figure 1. The flow chart for extracting features of MFCCs.

(1) Windowing: Sound changes over a wide range, and there is no fixed features operator for processing, so each frame is multiplied by the window function to increase the continuity of the left and right ends of the frame. Assume that the framed signal is \( x(n) \), \( n=0,1,...,N-1 \), \( N \) is the number of frames in a sample.

\[ x'(n) = x(n)w(n) \]  

(3)
Where \( w(n) \) is the Hamming window:

\[
w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1
\]  

(4)

(2) Fast Fourier Transform (FFT): The tool for extracting discrete frequency spectrum information from a discrete signal (sampled signal) is the FFT. After performing FFT on each frame of signal \( x'(n) \), the discrete power spectrum \( X(k) \) is obtained by modulo.

(3) Mel filtering: We divide the spectrum into several Mel filter banks according to human ear sensitivity. When calculating the MFCCs, the FFT spectrum can be converted into the Mel spectrum by passing through a set of Mel filter banks. The frequency response of each triangular filter is

\[
H_m(k) = \begin{cases} 
0 & k < f(m-1) \text{ or } k > f(m+1) \\
\frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k \leq f(m) \\
\frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \leq f(m+1)
\end{cases}
\]

(5)

Where \( f(m) \) is the center frequency and \( m = 1, 2, \ldots, M \), \( M \) is the number of Mel filters. The center frequency is converted from Mel scale.

(4) Logarithm and Discrete Cosine Transform (DCT): Calculating the logarithmic energy of the output of each filter bank and then doing a discrete cosine transform to obtain a set of cepstral coefficients:

\[
m_i = \log[X(k)H_m(k)], i = 1,2,\ldots,M
\]

(6)

\[
c_n = \sqrt{\frac{2}{M}} \sum_{i=1}^{M} m_i \cos\left(\frac{\pi n}{M} (i-0.5)\right)
\]

(7)

\( M \) is the number of filters, \( n = 1, 2, \ldots, L \), \( L \) is the number of cepstral coefficients, and \( c_n \) is the MFCCs per frame. The 12 MFCC parameters of each frame are taken as one column of the matrix, and the matrix formed in order is the MFCC feature matrix.

2.2.2. Further learning of features using Autoencoder

Autoencoder is essentially a 3-layer BP neural network, except that its output is equal to the input. In other words, the following relationship exists between the output layer and the input layer of Autoencoder:

\[
\hat{x} \approx x
\]

(8)

\( x \) is the input vector and \( \hat{x} \) is the output vector, called the reconstructed input vector.

Autoencoder is mainly composed of two parts: encoding process and decoding process. The specific process is as follows:

(1) Encoding process:

\[
h = f(W^{(i)}x + b^{(i)})
\]

(9)

Where \( f(\cdot) \) is the activation function. \( W^{(i)} \) represents the coding weight matrix and \( b^{(i)} \) represents a vector of bias parameter.

(2) Decoding process:
\[ \hat{x} = f(W^{(2)}h + b^{(2)}) \]  

(10)

Where \( W^{(2)} \) represents the decoding weight matrix and \( b^{(2)} \) represents a vector of bias parameter. The equation of the error at this time is as shown in equation (11).

\[ \text{loss} = \frac{1}{m} \sum_{i=1}^{m} (\hat{x} - x)^2 \]  

(11)

The error is back propagated. By adjusting the weights and biases, the error decreases along the gradient. Repeating learning and training until determining the weight and bias of each layer is corresponding to the minimum error, then stop training. At this point \( h \) removes the redundant input information and captures the most important factors that can represent the input data.

2.3 Classification and voting strategy

The features are input into a specific classifier. After classifying each frame of an audio sequence, then using a voting strategy to determine the final category. The idea of voting strategy is: The class ID is determined based on weak results voting, i.e. according to the principle that the minority is subordinate to the majority, which can be used to determine which category the sound sample belongs to. In this way, multiple weak classifiers are combined into one strong classifier, greatly improving the accuracy of classification [9].

3. Simulation experiment and result analysis

3.1. Sound data set

The sound data used in the experiment included sound of sand mining collected at different locations, sound of train running near the sand mining boat, sound of the engine, sound of flowing water, and sound of birds and wind. All the sounds are mixed with noise. These sounds were collected on the banks of the Yangtze River, where the number of no-sand-mining sounds is significantly less than that of sand mining.

3.2. Feature analysis

In order to directly observe the advantages of combined features, an audio file was randomly selected from sand mining sounds. The extracted MFCCs are drawn in figure 2(a), the features extracted by the combination of MFCC abstractor and Autoencoder are shown in figure 2(b). Where the coefficient length of each frame of MFCCs is 12; The number of neurons in the input and output layers of the Autoencoder is 12, the number of neurons in the hidden layer is 8, the activation function of the hidden layer is tanh and the activation function of the output layer is softmax.

![Figure 2. Features. (a) MFCCs, (b) The combination of MFCCs and Autoencoder](image-url)
As can be seen from figure 2(a), in the first 6 dimensional features, the variations of each frame of the 249 frames are roughly the same, but in the latter 6 dimensional features, although the variation of the majority of the frames is uniform, the variation of the minority of the frames is very confusing. This shows that there is indeed a small part of noise in the MFCCs. But it can be seen from figure 2(b) that the feature changes of each dimension of the 249 frames are consistent, and the first 6 dimensional changes are also more uniform than those of the first 6 dimensional changes in figure 2(a).

By comparing the two figures, we can see that features extracted using the MFCC abstractor and Autoencoder can indeed filter out most of the noise and better express the features of sand mining sound.

3.3. Perform experiments and results

After learning the features using MFCC abstractor and Autoencoder, the features are input into SVM, KNN and BP neural network algorithms respectively. At the same time, in order to compare the advantages of the combined features, the features extracted only using MFCC abstractor are input into the same classifiers. In this paper, the kernel of SVM is Radial Basis Function, k=1 in KNN, and BP neural network is a 3-layer network structure. The specific results are shown in table 1, table 2, and table 3.

Table 1. Results of multi-classification using SVM.

| Sound type               | MFCC+ Autoencoder + SVM | MFCC + SVM  |
|--------------------------|-------------------------|-------------|
| Sound of sand mining     | 100%                    | 95.93%      |
| Sound of train running   | 97.14%                  | 100%        |
| Sound of the engine      | 100%                    | 100%        |
| Sound of flowing water   | 82.98%                  | 46.81%      |
| Sound of birds and wind  | 87.50%                  | 79.17%      |

Table 2. Results of multi-classification using KNN.

| Sound type               | MFCC+ Autoencoder + KNN | MFCC + KNN  |
|--------------------------|-------------------------|-------------|
| Sound of sand mining     | 100%                    | 100%        |
| Sound of train running   | 91.43%                  | 91.43%      |
| Sound of the engine      | 100%                    | 73.91%      |
| Sound of flowing water   | 93.62%                  | 68.09%      |
| Sound of birds and wind  | 91.67%                  | 91.67%      |

Table 3. Results of multi-classification using BP neural network.

| Sound type               | MFCC+ Autoencoder + BPNN | MFCC + BPNN |
|--------------------------|--------------------------|-------------|
| Sound of sand mining     | 100%                    | 100%        |
| Sound of train running   | 88.57%                  | 80%         |
| Sound of the engine      | 100%                    | 91.30%      |
| Sound of flowing water   | 95.74%                  | 85.11%      |
| Sound of birds and wind  | 95.83%                  | 91.67%      |

As seen in table 1, table 2 and table 3, the experimental results of combined features are mostly better than the experimental results of features extracted by the MFCC abstractor. The results show that the combined features can greatly improve the accuracy of sound recognition compared with the features extracted only by MFCC.

If only two categories are considered, that is, only the “Sound of sand mining” and “Not sound of sand mining” are judged. The specific results are shown in table 4, table 5, and table 6.
Table 4. Results of binary-classification using SVM.

| Sound type              | MFCC+ Autoencoder + SVM | MFCC + SVM |
|-------------------------|-------------------------|------------|
| Sound of sand mining    | 100%                    | 100%       |
| Not Sound of sand mining| 72.87%                  | 59.69%     |

Table 5. Results of binary-classification using KNN.

| Sound type              | MFCC+ Autoencoder + KNN | MFCC + KNN |
|-------------------------|-------------------------|------------|
| Sound of sand mining    | 100%                    | 99.59%     |
| Not Sound of sand mining| 92.25%                  | 65.12%     |

Table 6. Results of binary-classification using BP neural network.

| Sound type              | MFCC+ Autoencoder + BPNN | MFCC + BPNN |
|-------------------------|--------------------------|-------------|
| Sound of sand mining    | 100%                     | 100%        |
| Not Sound of sand mining| 99.22%                   | 79.84%      |

It can be seen from table 4, table 5, and table 6 that among the three algorithms, the results of experiments using combined features to identify each type of sound are best.

It can also be seen from the above six tables that the accuracy of identification of sand mining sounds using the combined features is all 100%, which indicates that the combined features are particularly suitable for the identification of sand mining sounds, and it can also perfectly realize the task of monitoring sand mining.

Figure 3 below shows the average experimental accuracy of multi-classification and binary-classification.

![Figure 3. Average experimental accuracy.](image)

It can be seen from figure 3 that the average experimental accuracy of multi-classification and binary-classification have been greatly improved by using the combined features, whether the classifier is SVM, KNN or BP neural network. It shows that this way of extracting features by combination of MFCC abstractor and Autoencoder can indeed improve the ability to express features and better accomplish the task of monitoring sand mining.

At the same time, it can be clearly seen from the results that whether it is binary-classification or multi-classification, regardless of which classifier is adopted, the average experimental accuracy is above 86%. That's because every method of classification uses the idea of voting strategy. It is illustrated that multiple weak classifiers are combined into a strong classifier, in which each weak classifier votes,
and the minority is subordinate to the majority, returning the result of the final vote, so the accuracy of classification is indeed very high.

4. Conclusion

In this paper, an approach for monitoring sand mining based on sound is presented. Different from the method of extracting acoustic features by using MFCC abstractor, this paper uses a combination of MFCC abstractor and Autoencoder to extract acoustic features and a voting strategy is used to determine the category of sound, so as to judge whether sand mining activities have occurred. The above experiments proved that the features extracted by combination of MFCC abstractor and Autoencoder can well express the sand mining sounds and the various sounds associated with them, and also proved that the voting strategy based on combination of features can achieve good results in sand mining sound recognition. It is believed that the sand mining monitoring system with this artificial intelligence method can play a very good role in the sand mining management of the river. The next study of us is to apply this combination of features to more types of recognition of environmental sound, so that we can better use the information in natural environment sounds to serve humans.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 51605061), the Fundamental Science and Frontier Technology Research Foundation of Chongqing, Chongqing, China. (cstc2017jcyjA1592 and cstc2017jcyjA1475), and project of Chongqing Huize Technology Development Co., Ltd.(2018500015000110).

References

[1] Zahorian S, Guzewich P, Chen X and Zhang H 2017 The relative importance of static versus spectral change acoustic features for automatic speaker identification, J. Acoust. Soc. Am. 141 (5) 3915-3915.
[2] Mrquez-Molina M and Surez-Guerra S 2014 Aircraft take-off noises classification based on human auditory’s matched features extraction, Appl Acoust. 84 83-90.
[3] Jensen J and Tan Z H 2015 Minimum Mean-Square Error Estimation of Mel-Frequency Cepstral Features–A Theoretically Consistent Approach, IEEE Trans. Audio Speech Lang. Process. 23 (1) 186-197.
[4] Cao J, Zhao T, Wang J, Wang R and Chen Y 2017 Excavation equipment classification based on improved MFCC features and ELM, Neurocomputing 261 231-241.
[5] Liu J H, Zheng W Q and Zou Y X 2015 A Robust Acoustic Feature Extraction Approach Based on Stacked Denoising Autoencoder, in: IEEE International Conference on Multimedia Big Data, pp 124-127.
[6] Nishimura Y, Shinozaki T, Iwano K and Furui S 2004 Noise-robust speech recognition using multiband spectral features, J. Acoust. Soc. Am. 116 (4) 2480-2480.
[7] Feixiang Z, Yongxiang L, Kai H, Shuanghui Z, and Zhongshuai Z (2018). Radar hrrp target recognition based on stacked autoencoder and extreme learning machine. Sensors 18(2) 173-187
[8] Zhou C and Paffenroth R C 2017 Anomaly Detection with Robust Deep Autoencoders, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 665-674. doi:10.1145/3097983.3098052.
[9] Tamvakis A, Tsirtsis G, Niros A D and Spatharis S 2018 Optimized Classification Predictions with a New Index Combining Machine Learning Algorithms. Int. J. Artif. Intell. Tools. 2