Research on Sound Source Recognition Algorithm of Pickup Array Based on Adaptive Background Noise Removal

Chengyu Hou, Liu Can, and Di Chen(✉)

Harbin Institute of Technology, Harbin, China
dchen@hit.edu.cn

Abstract. Nowadays, the pickup array is used in a large number of occasions, such as human voice recognition, audio conference, video conference and sound source localization. The research of sound source recognition algorithm based on pickup array has broad application prospects in the military field. The sound source recognition technology at this stage is implemented by a relatively fixed pickup array. However, due to the high requirements for the number of array elements, it faces severe environmental noise interference. Therefore, the sound source signal needs to be pre-processed before being formally processed. This paper discusses the sound source recognition algorithm based on the pickup array, which reduces the influence of environmental noise interference by preprocessing the sound source signal; realizes the target sound source recognition through feature extraction and the establishment of a recognition model. This article starts with the study of the preprocessing method of the sound source signal of the L-shaped pickup array node, and discusses an LMS noise cancellation model based on an improved variable step size. At the same time, this article identifies the target sound source signal and uses the MFCC feature extraction method. On the basis, the MFCC feature extraction method for high frequency suppression is given, and then the sound source recognition algorithm based on GMM-UBM is introduced.

Keywords: L-shaped pickup array · Noise cancellation · Sound source identification

1 Introduction

Since the array signal processing technology was successfully introduced into the field of speech signal processing by Professor Silverman and others, the use of pickup arrays in speech signal processing has become a new research hotspot [1]. Nowadays, pickup arrays are used in many occasions, such as human voice recognition, video conferencing, sound source localization, etc. [2–4]. However, the sound source identification research based on the pickup array is rarely seen in the military field due to its low anti-interference ability and complex terrain environment. However, due to the relatively long sound wave wavelength, its unique diffraction characteristics and low-cost low-power consumption. The pickup makes a very high economic benefit [3].

X. Wang et al. (Eds.): AICON 2021, LNICST 396, pp. 346–361, 2021.
https://doi.org/10.1007/978-3-030-90196-7_29

At this stage, the sound source identification is realized by a relatively fixed pickup array. However, due to the high requirement on the number of array elements, it faces severe environmental noise interference [4–7]. Therefore, the sound source signal needs to be preprocessed first. At the same time, due to the sound signal It is a wideband signal, and the phase difference output after it is received is not only related to the direction, but also related to the signal frequency, which increases the amount of calculation for the sound source recognition algorithm.

The main research content of this paper is the sound source recognition algorithm based on the L-shaped pickup array, which reduces the influence of environmental noise interference by preprocessing the sound source signal; realizes the target sound source recognition through feature extraction and establishment of a recognition model [8–10].

2 Adaptive Variable Step Size NLMS Algorithm

2.1 Analysis of L-Shaped Pickup Array Structure

Figure 1 shows the signal model of the L-shaped pickup array. Since the pickup array needs to be a two-dimensional or three-dimensional structure, the algorithm of the latter is more complicated and costly. Therefore, the design of the pickup array structure in this paper is L-shaped.

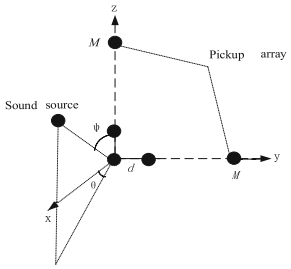


Fig. 1. Signal model of L-shaped pickup array

The sound sources used in this paper are tank sound, truck sound and infantry walking sound. The noise of the three sound sources is between 10 Hz and 850 Hz. Therefore, the half-wavelength theory can be used to obtain the expression of the distance d of the pickup array:

$$d \leq \frac{1}{2}\lambda = \frac{1}{2} \times \frac{c}{f} = \frac{340 \text{ m/s}}{2 \times 850 \text{ Hz}} \approx 0.20 \text{ m} \tag{1}$$

In formula (1), d is the element spacing between two adjacent elements in the L-shaped array, c represents the propagation speed of sound in the air (under a standard atmospheric pressure, the speed is 340 m/s), and λ is the wavelength of the sound wave, f is the sound source frequency.

From formula (1), it can be seen that if the array element spacing $d \leq 0.2$, the aperture of the array is no longer limited by the half-wavelength theory, so in this article d = 0.20 m; Considering the computational complexity and cost issues, choose to install 3 pickup elements in the horizontal and vertical directions, that is, the horizontal and vertical directions share one element at the junction. Considering that environmental noise will affect the reception of the target signal by the pickup array in the actual situation, the next section will discuss the preprocessing process of the sound source signal received by the pickup array, and uses the method of noise cancellation to restore the sound spectrum structure of the sound source.

2.2 Nodal Sound Source Signal Preprocessing

2.2.1 Analysis of Adaptive Noise Cancellation System

The adaptive noise cancellation system of the pickup element is shown in Fig. 2. In order to obtain the environmental background noise, this article uses two types of high and low sensitivity pickups. The high-sensitivity pickup mainly collects the mixed signal of the sound source and the noise, namely: $x_i(t) = s(t) + n_i(t)$ A low-sensitivity pickup is placed on the top of the pickup array to collect background noise $n_0(t)$ that is not related to the sound source signal $s(t)$ but related to $n_i(t)$ in the experimental environment. After the background noise $n_0(t)$ passes through the adaptive filter, a noise estimation signal $\hat{n}_i(t)$ can be obtained, which is subtracted from the main signal $\hat{n}_i(t)$ can get the required sound source signal after denoising, namely:

$$y_i(k) = x_i(t) - \hat{n}_i(t) = s(t) + n_i(t) - \hat{n}_i(t) \tag{2}$$

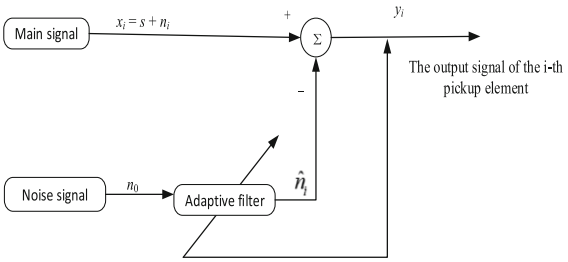


Fig. 2. Adaptive noise cancellation system

Figure 3 is a schematic diagram of the adaptive noise cancellation algorithm. The pollution signal is $x_i(k)$ in Fig. 2, namely:

$$x_i(k) = s(k) + n_i(k) \tag{3}$$

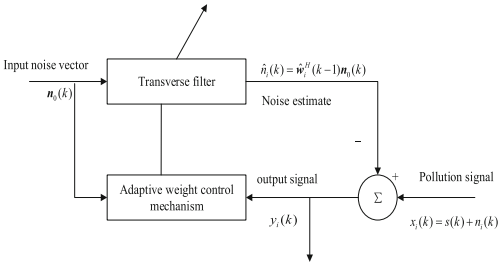


Fig. 3. Adaptive noise cancellation algorithm

In Eq. (3), i represents the i-th high-sensitivity pickup, and $\boldsymbol{n}_0(k)$ is the noise vector collected by the low-sensitivity pickup. A noise estimate $\hat{n}_i(k)$ can be obtained through the transversal filter in the figure, namely:

$$\hat{n}_i(k) = \hat{\boldsymbol{w}}_i^H(k-1)\boldsymbol{n}_0(k) \tag{4}$$

The $\hat{\boldsymbol{w}}_i(k-1)$ in Eq. (4) represents the least mean square estimation of the weight vector of the transversal filter in the system at k–1.

From Eq. (3), the output signal $y_i(k)$ can be:

$$y_i(k) = s(k) + n_i(k) - \hat{n}_i(k) \tag{5}$$

Then its mean square value is:

$$E[y_i^2(k)] = E[s^2(k)] + E[(n_i(k) - \hat{n}_i(k))^2] + 2E[s(k)(n_i(k) - \hat{n}_i(k))] \tag{6}$$

It can be obtained by the irrelevant nature of the signal and noise:

$$E[y_i^2(k)] = E[s^2(k)] + E[(n_i(k) - \hat{n}_i(k))^2] \tag{7}$$

It can be seen that in the case of the minimum mean square of the transversal filter, $\hat{n}_i(k)$ and $\boldsymbol{n}_0(k)$ are the closest, then the output signal $y_i(k)$ and the target sound source signal $s(k)$ are also the closest at this time, and the original signal can be restored to the maximum extent.

2.2.2 Adaptive Noise Cancellation Algorithm

The most classic type of adaptive noise cancellation algorithms is the LMS algorithm, namely:

$$e(k) = d(k) - \hat{\mathrm{w}}^H(k-1)\mathrm{x}(k) \tag{8}$$

$$\hat{\mathrm{w}}(k) = \hat{\mathrm{w}}(k-1) + \mu\mathrm{x}(k)e^*(k) \tag{9}$$

In formula (9), μ is the step factor, which is a fixed value, and corresponds to the above:

$$d(k) = s(k), \hat{w}(k) = \hat{w}_i(k), u(k) = x_i(k) \tag{10}$$

The convergence conditions of the algorithm are:

$$0 < \mu < \frac{2}{\lambda_{\max}} \tag{11}$$

In formula (11), $\lambda_{\max}$ represents the largest eigenvalue corresponding to the autocorrelation matrix of the input signal $\mu(k)$.

LMS adaptive calculation cannot achieve the best compromise between convergence rate and steady-state error. As μ increases, the rate of convergence is greater, but at the same time the steady-state error will be greater. In order to solve this contradiction, the step factor μ can be adjusted with the iteration process.

The normalized least mean square (NLMS) adaptive algorithm is proposed on this basis, and its convergence result has nothing to do with the strength of the input signal, so its step adjustment function is as follows:

$$\mu(k) = \frac{\tilde{\mu}}{\delta + \|\mathrm{u}(k)\|^2} \tag{12}$$

In formula (12), $\tilde{\mu}$ is an adaptive constant; δ is a small constant greater than 0, which is used to solve the calculation problem with a denominator of 0.

The NLMS algorithm can be regarded as a variable step size algorithm. Its convergence rate is faster than that of the LMS algorithm, but it cannot effectively solve its contradiction with steady-state errors.

Another idea that can effectively solve this problem is to use a larger step size to increase the convergence rate at the beginning of the algorithm iteration, and use a smaller step size when the iteration is about to complete to reduce the steady-state error. The change of $u(k)$ is related to the error signal $e(k)$. Professor Qun Niu developed a variable step size LMS algorithm in 2018, and the $u(k)$ is:

$$\mu(k) = \beta[1 - \exp(-\alpha \left| e^2(k)e(k-1) \right|)] \tag{13}$$

In formula (13), both α and β are constants.

The form of formula (13) is relatively simple, and the step length changes slowly when the error approaches 0, which optimizes the convergence characteristics, but does not describe the physical meaning of the exponential term. The adaptive step size NLMS algorithm proposed in this paper takes into account the interference of colored noise and uses the third-order correlation of the error signal $e(k)$ to adjust the step size, which can effectively improve the contradiction between the convergence speed and the steady-state error. Long and unaffected by system noise, its $u(k)$ is:

$$\mu(k) = \beta[1 - \exp(-\alpha |e(k)e(k-1)e(k-2)|)] \tag{14}$$

In formula (14), $a > 0$, the function is to control the step change trend of the adaptive algorithm in the iterative process, $0 < \beta < 2/\lambda_{\max}$, is a constant used to control the step change interval. When a is constant, the initial step size and convergence rate increase with the increase of β, but the steady-state error will also increase; when β is constant, the change of step size tends to be flat with the increase of a, and the rate will decrease during the convergence process. Increase, but the steady-state error also increases. Therefore, the algorithm in this paper can choose a smaller a value and a larger β value, which can effectively improve the convergence rate and steady-state error.

2.2.3 Experimental Results and Analysis

Based on the content of the previous section, this section compares the three algorithms of LMS, NLMS and the algorithm in this paper, and analyzes the recovery and convergence performance of the signal added with white noise/color noise. Table 1 below is the white noise condition experimental simulation parameters.

Table 1. Simulation experiment parameters of various adaptive algorithms under white/color noise conditions

Algorithm	SNR	Filter order	Fixed step μ	Adaptive constan $\tilde{\mu}$	Constant α	Constant β
LMS	15 dB	10	0.005	—	—	—
NLMS	15 dB	10	—	0.05	—	—
Algorithm	15 dB	10	—	—	1000	0.08

The experimental results are shown in the figure below:

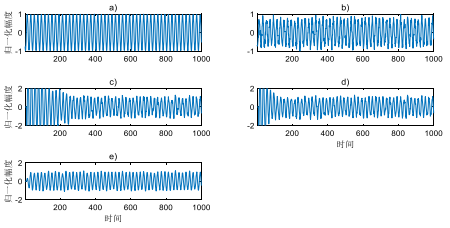


Fig. 4. Time-domain types of algorithms under white noise under white noise conditions

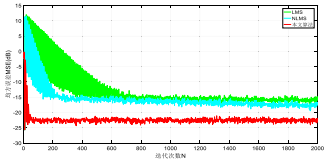


Fig. 5. The learning curve of the three types of algorithms under white noise conditions

Figure 4 is a comparison diagram of the time-domain signals output by the three types of algorithms under white noise conditions, where a) is the original signal $\sin(0.1\pi t + 10)$, b) is the signal with white noise, and c) is the LMS cancellation After the output signal, d) is the output signal after NLMS cancellation, e) is the output signal after cancellation by the algorithm in this paper. It can be seen from the figure that the noise-added signal has some residual noise after passing through the LMS algorithm and the

NLMS algorithm. The algorithm in this paper has a significantly better effect of filtering noise.

Figure 5 is the learning curve of the three types of algorithms under white noise conditions. The number of sampling points is 2000, and the simulation iterations are 150 times. From the figure, it can be seen that the LMS algorithm completes convergence after about 800 iterations, while the NLMS algorithm is 300 times. The algorithm used in this article Convergence is reached after only 50 iterations. At the same time, it can be seen from the figure that the steady-state error after convergence is the last one. It can be seen that the algorithm in this paper has a faster convergence rate compared with the two algorithms of LMS and NLMS. As well as lower steady-state error, it can deal with the contradiction between convergence rate and steady-state error very effectively.

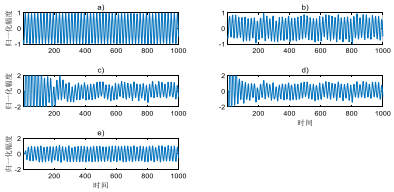


Fig. 6. Three types of algorithms output time-domain signals under colored noise conditions

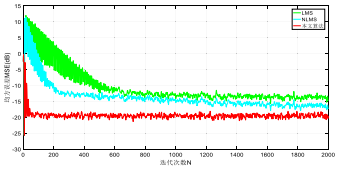


Fig. 7. The learning curve of the three types of algorithms under the color noise condition

Figure 6 is a comparison diagram of the time-domain signal output by the three types of algorithms under the color noise condition, where a) is the original signal sin $(0.1\pi t + 10)$, b) is the signal with white noise, and c) is the LMS cancellation After the output signal, d) is the output signal after NLMS cancellation, e) is the output signal after cancellation by the algorithm in this paper. It can be seen from the figure that the color-added noise signal has more noise residue and distortion after passing through the LMS algorithm and the NLMS algorithm. The algorithm in this paper has a significantly better effect of filtering noise without distortion.

Figure 7 is the learning curve of the three types of algorithms under the color noise condition. The number of sampling points is 2000, and the simulation iteration is 150 times. It can be seen from the figure that the LMS algorithm has completed convergence after about 800 iterations, while the NLMS algorithm is 400 times. The algorithm used in this article Convergence is reached after only 60 iterations; at the same time, it can be seen from the figure that the steady-state error after convergence is the last one. It can be seen that the algorithm in this paper has a faster convergence rate compared with LMS and NLMS As well as lower steady-state error, it can deal with the contradiction between convergence rate and steady-state error very effectively.

In order to verify the effect of this algorithm in practical applications, a section of tank traveling sound is used as an experiment. Considering the serious influence of colored noise on the spectrum structure of the sound source in the actual environment, the additional frequencies of the original sound source are 1000 Hz, 1500 Hz, 2000 Hz and The sound of 2500 Hz is used to simulate the interference of colored noise. The specific experimental parameters are as follows (Table 2).

Table 2. Experimental parameters of adaptive noise reduction spectrum of actual tank sound source

Filter order	Sampling frequency	Step change trend control constant α	Step change interval control constant β
100	48 kHz	1	0.01

The experimental results are as follows:

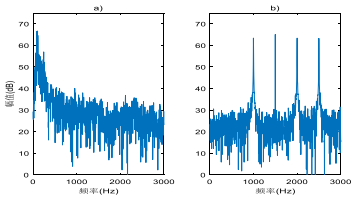


Fig. 8. a) Comparison of the original sound spectrum and b) the colored interference sound spectrum of the low-sensitivity channel.

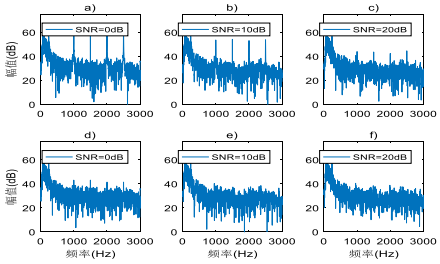


Fig. 9. Comparison of adaptive noise vs. anechoic spectrum under different SNR conditions of high-sensitivity channels

Figure 8 shows the original sound spectrum a) of the tank moving sound source, and the color noise interference sound spectrum obtained by the low-sensitivity pickups in the array b), while Fig. 9 shows the noise-containing spectrum obtained by the high-sensitivity channel and preprocessed Sound spectrum comparison, where a), b), and c) are the colored noise spectrum under different signal-to-noise ratio conditions, and d), e), and f) are the corresponding adaptive noise cancellation spectra. From the above results,

it can be seen that the algorithm in this paper can effectively filter the interference of colored noise, and at the same time SNR will affect its removal effect. There is some noise interference under the condition of 0 dB, and the effect of 10 dB and 20 dB is better.

3 Sound Source Recognition Algorithm Based on L-Shaped Pickup Array

3.1 MFCC Feature Extraction Method for High Frequency Suppression

The most commonly used feature extraction method for the target sound source is the MFCC feature extraction method, which is to obtain the characteristic parameters of the target sound source by simulating the non-linear mapping characteristics of the human ear when receiving sound. From the previous analysis, it can be seen that the spectral characteristics of the target sound source are concentrated in the low-frequency region, so the target sound source needs to be subjected to low-pass filtering before feature extraction to suppress high-frequency noise. The specific process is as follows.

Fig. 10. MFCC feature extraction flowchart

Figure 10 shows the MFCC feature extraction process. The target sound source signal is first passed through a low-pass filter to suppress the high frequency part, and then the signal is processed by FFT after framing and windowing, so as to change the signal spectrum into a linear spectrum, and then use mei After the Mf filter is processed by the logarithmic transformation, the logarithmic nonlinear spectrum after dynamic range compression can be obtained, where the mel frequency is $f_{mel} = 2595 \ \log_{10}(1+f/700)$, and the mel frequency filtering is achieved by Mf triangular bandpass filters. The sound source is converted from a linear frequency spectrum to a mel frequency spectrum. The transfer function is as follows:

$$H_i(k) = \begin{cases} 0 & (k < f(i-1)) \\ \frac{k-f(i-1)}{f(i)-f(i-1)} & (f(i-1) \le k \le f(i)) \\ \frac{f(i+1)-k}{f(i+1)-f(i)} & (f(i) \le k \le f(i+1)) \\ 0 & (k > f(i+1)) \end{cases} \tag{15}$$

Finally, perform discrete cosine transform on it, namely:

$$o(l) = \sum_{i=1}^{M_f} S(i) cos\left(\frac{l\pi(i+1/2)}{M_f}\right) 1 \le l \le D/2 \tag{16}$$

$S(i)$ in Eq. (16) is the logarithmic spectrum obtained by the i-th triangular filter, and $o(l)$ is the l-dimensional static characteristic of the obtained target sound source. Then,

the difference operation with an interval of 2 can be performed on different frames to obtain the dynamic characteristics. The combination of the two forms a D-dimensional MFCC feature vector, which is the MFCC feature parameter of the target sound source signal.

3.2 GMM-UBM Sound Source Recognition Algorithm

GMM model is mainly used to recognize the speaker's voice. Its advantage is that it does not need to care about the semantic and contextual connection during training and pattern matching, so it is suitable for sound source recognition.The disadvantage is that its parameter scale and characterization ability are not coordinated. When the decomposed Gaussian component is relatively small, the accuracy of the characteristic model obtained is low. Therefore, it needs to be improved. GMM-UBM algorithm is improved on this basis, and its target sound source identification process is shown in the figure below (Fig. 11).

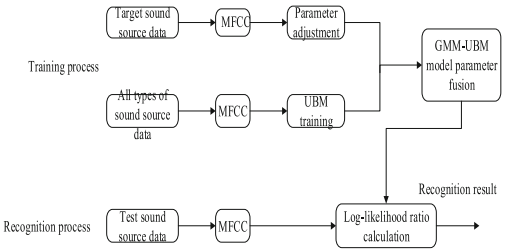


Fig. 11. Sound source recognition process of GMM-UBM model

First, a large number of sound source data need to be UBM training to obtain its model parameter h_{UBM} Then, based on the maximum posterior probability adaptive principle, the model parameter H is obtained by fine-tuning the data obtained from a small amount of target sound $h_{t\arg et}$ and the two model parameters were fused to obtain the parameter h of GMM-UBM $h_{fus,s}$. When identifying the target sound source, it is necessary to first extract the features of the identified sound source data, then calculate the posterior probability of UBM and GMM-UBM model, and then calculate the logarithmic likelihood ratio of the two models respectively for scoring and identification.

3.2.1 GMM Model

Suppose that the sound source feature vector of the t-th frame of the target sound source signal obtained after MFCC feature extraction is $o_t = [o_t(1), o_t(2)...o_t(D)]$, and the likelihood function of its GMM model is fitted with G Gaussian components as follows:

$$p(\mathrm{o}_t|h) = \sum_{i=1}^{G} \omega_i p_i(\mathrm{o}_t) \tag{17}$$

$$p_i(\boldsymbol{o}_t) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \mu_i)(\boldsymbol{\Sigma}_i)^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)\prime\right] \tag{18}$$

$$\sum_{i=1}^{G} \omega_i = 1 \tag{19}$$

In formula (17), $p(o_t|h)$ is the Gaussian mixture function of ot; h is the parameter set of the GMM model; $p_i(o_t)$ is the distribution function of the i-th Gaussian component, and ω_i is the corresponding weight. In Eq. (18), μ_i is the mean vector of the Gaussian component distribution function, and Σ_i is the corresponding covariance matrix (usually an DxD dimensional diagonal matrix).

The principle of GMM training is: Given a training sample, use the Expectation Maximum (EM) method to obtain the maximum likelihood estimation of h. The likelihood of the model parameter set h of the T-frame sound training sample is calculated as follows:

$$L(h|\mathrm{O}) = P(\mathrm{O}|h) = \prod_{t=1}^{T} p(\mathrm{o}_t|h) \tag{20}$$

From Eq. (20), the maximum likelihood of h can be estimated as follows:

$$\hat{h} = \arg\max_{\hat{h}} L(h|\mathrm{O}) = \arg\max_{\hat{h}} P(\mathrm{O}|h) = \arg\max_{\hat{h}} \prod_{t=1}^{T} p(\mathrm{o}_t|h) \tag{21}$$

When the initial value $h_0 = \{(\omega_i^{(0)}, \mu_i^{(0)}, \Sigma_i^{(0)})\}$ A of h is given, the EM method can be used to loop iteratively to obtain its maximum likelihood estimation solution.

3.2.2 UBM Model

Affected by the number of Gaussian components, the recognition performance of the GMM model is also related to it. The greater the number of Gaussian components, the better the recognition effect. However, with the increase of Gaussian components, the corresponding target sound source data required increases, which leads to the increase of model parameters that need to be estimated, and the amount of calculation is huge. Based on this, a UBM training algorithm is proposed. The principle of the UBM algorithm is to use the EM algorithm to train all types of sound source samples to obtain a GMM model that is not related to the sound source type. This model is the feature model $h_{UBM} = \{(\omega_{UBM,i}, \mu_{UBM,i}, \Sigma_{UBM,i})\}$ common to all types of sound sources. After obtaining the required GMM model, based on the shared feature model, only a few target sound sources can be adapted to the model parameters based on the maximum posterior probability criterion. The process includes parameter fine-tuning and parameter fusion based on UBM.

If the characteristic sample of the sound source to be identified is $O_{t\arg et} = \{O_{t\arg et,1}, O_{t\arg et,2}..., O_{t\arg et,T}\}$, the posterior probability of the j Gaussian components of the parameter adjustment model is as follows:

$$Pr\left(j|\mathrm{o}_{target,t}, h_{UBM}\right) = \frac{\omega_{UBM,j} p_j\left(\mathrm{o}_{target,t}, \mu_{UBM,j}, \Sigma_{UBM,j}\right)}{\sum_{i=1}^{G} \omega_{UBM,i} p_i\left(\mathrm{o}_{target,t}, \mu_{UBM,i}, \Sigma_{UBM,i}\right)} \tag{22}$$

The weights are as follows:

$$\omega_{target,j} = \frac{\sum_{t=1}^{T} Pr\left(j|o_{target,t},\ h_{UBM}\right)}{T} \tag{23}$$

The mean vector is as follows:

$$\mu_{target,j} = \frac{\sum_{t=1}^{T} Pr\left(j|o_{target,t},\ h_{UBM}\right)o_{target,t}}{\sum_{t=1}^{T} Pr\left(j|o_{target,t},\ h_{UBM}\right)} \tag{24}$$

The covariance matrix is as follows:

$$\boldsymbol{\Sigma}_{target,j} = \frac{\sum_{t=1}^{T} Pr\left(j|o_{target,t},\ h_{UBM}\right)\left(\boldsymbol{o}_{target,t} - \mu_{target,j}\right)'\left(\boldsymbol{o}_{target,t} - \mu_{target,j}\right)}{\sum_{t=1}^{T} Pr\left(j|\boldsymbol{o}_{target,t},\ h_{UBM}\right)} \tag{25}$$

After the above calculation is completed, the UBM parameter set and the fine-tuned parameter set are combined to obtain the GMM-UBM model $h_{fus} = \{(\omega_{fus,i},\ \mu_{fus,i},\ \Sigma_{fus,i})\}$ of the target sound source to be identified, where

$$\omega_{fus,j} = \alpha_j^{\omega}\omega_{target,j} + \left(1 - \alpha_j^{\omega}\right)\omega_{UBM,j} \tag{26}$$

$$\mu_{fus,j} = \alpha_j^{\mu}\mu_{target,j} + \left(1 - \alpha_j^{\mu}\right)\mu_{UBM,j} \tag{27}$$

$$\boldsymbol{\Sigma}_{fus,j} = \alpha_j^{\boldsymbol{Sigma}}\Sigma_{target,j} + \left(1 - \alpha_j^{\boldsymbol{Sigma}}\right)\left(\boldsymbol{Sigma}_{UBM,j} + \boldsymbol{mu}_{target,j}^2\right) - \boldsymbol{mu}_{fus,j}^2 \tag{28}$$

$$\alpha_j^{\rho} = \left\{\alpha_j^{\omega},\ \alpha_j^{\mu},\ \alpha_j^{\Sigma}\right\} = \frac{\sum_{t=1}^{T} Pr\left(j|o_{target,t},\ h_{UBM}\right)}{\sum_{t=1}^{T} Pr\left(j|o_{target,t},\ h_{UBM}\right) + \tau^{\rho}},\ \rho \in \{\omega, \mu, \Sigma\} \tag{29}$$

3.2.3 Voice Scoring Recognition

After training all the sound source data to obtain the GMM parameters, the sound source identification can be performed. You only need to calculate the likelihood function corresponding to the h_{target} G of the target sound source, and then traverse the maximum posterior probability to obtain the estimation of the maximum posterior probability. The recognition results are as follows:

$$\hat{s} = \arg\max_{1 \le s \le S} p\left(O|h_{fus,s}\right) = \arg\max_{1 \le s \le S} \sum_{t=1}^{T} \log p\left(o_t|h_{fus,s}\right) \tag{30}$$

In Eq. (30), $\hat{s}$ represents the recognition result of the sound source to be identified, and $h_{fus,s}$ is the GMM-UBM model parameter of the s-th sound source.

If there is only one type of sound source to be identified and the accuracy requirements are low, only this type of sound can be trained.Then calculate its log likelihood:

$$\Lambda(\mathrm{O}) = \frac{1}{T}\sum_{t=1}^{T} \log p\left(\mathrm{o}_t | h_{fus,s}\right) - \log p(\mathrm{o}_t | h_{UBM}) \tag{31}$$

In formula (31), h_{fus} is the GMM-UBM model parameter of the target sound source signal; h_{UBM} is the UBM model parameter of all types of sound. You can also reduce the amount of calculation by setting the decision threshold.

The larger $\Lambda(\mathrm{O})$ is, the greater the similarity of features between the sound source to be identified and the target sound source is, and the smaller the $\Lambda(\mathrm{O})$, the higher the similarity between the sound source to be identified and E that is not related to the target sound source.

3.3 Experimental Results and Analysis

First do MFCC feature extraction, and the simulation parameters are set as follows (Table 3):

Table 3. MFCC feature extraction experimental parameters

Sampling frequency	Feature dimension	Filter order	Framing number	Oneframe duration	Frame shift ratio	FFT	Cut-off frequency
48 kHz	24	24	100	20 ms	1/4	2048	2000 Hz

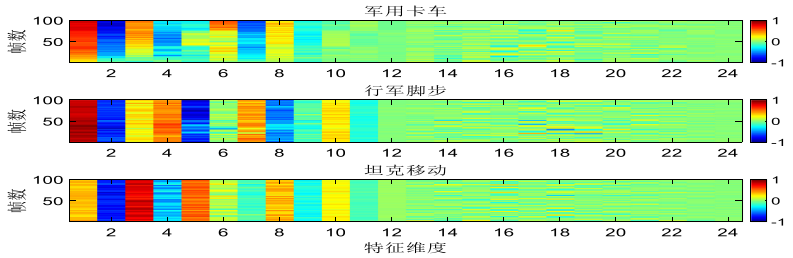
The experimental results are as follows (Fig. 12):

As shown in Fig. 13, it is a normalized feature map of MFCC feature extraction for three sound sources of military truck sound, marching sound and tank sound. It can be seen that if high frequency suppression is not done, the MFCC static characteristics of various sound sources The difference in the first three dimensions is very small, and the distinction of dynamic features is not good. After high-frequency suppression, the static features of the target sound source are very different, and it clearly reflects the difference between different frames. Therefore, it is necessary to perform high frequency suppression before performing MFCC feature extraction.

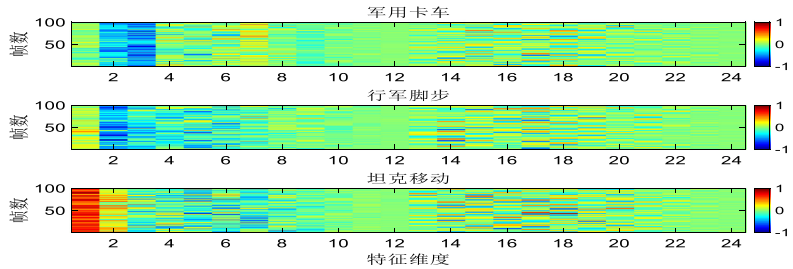
Then the target sound source is identified. The following table shows the experimental parameters of feature extraction of the GMM-UBM algorithm and the experimental parameters of its training and recognition process (Table 4).

The experimental results are as follows:

As shown in Fig. 14 above, the false alarm rate is 1.099%. Figure 14 is the scoring result of the normalized fusion of the two tank sounds of m109 and Leopard by the GMM-UBM algorithm, where m109 is the 11th type of sound source, and leopard is the 10th type. From the above figure, we can see that the algorithm is for non-targets. The score of the sound source category is lower, which means that when the normalized threshold is the same, the false alarm rate is smaller, that is, the algorithm has fewer incorrectly associated nodes.



a) MFCC feature map without high frequency suppression



b) MFCC feature map for high frequency suppression

Fig. 12. Comparison of MFCC features without high frequency suppression and high frequency suppression

Table 4. GMM-UBM feature extraction experimental parameters

Sampling frequency	Feature dimension	Filter order	Framing number	One frame duration	Frame shift ratio	FFT	Cut-off frequency
4819.98 kHz	24	24	550	24 ms	1/4	4096	2000 Hz

Table 5. GMM-UBM training and recognition experiment parameters

Experimental model	Number of test samples	Number of training samples	Gaussian component number	Number of samples per type	Model correlation factor
GMM-UBM	14	14	128	15	10

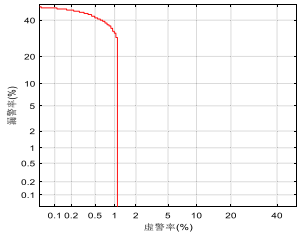
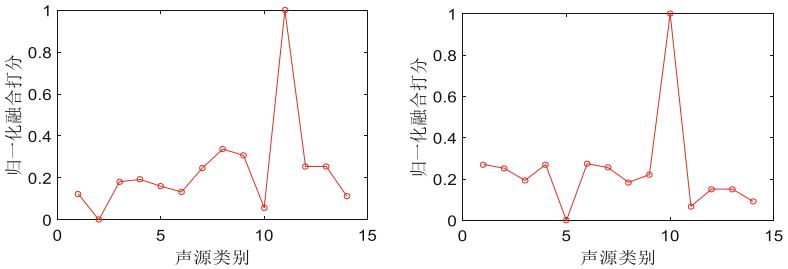


Fig. 13. GMM-UBM warning error curve



a) m109 tank sound score b) Leopard tank sound score

Fig. 14. Comparison of the scoring results of the two types of target sound sources by GMM-UBM

4 Conclusion

This paper mainly studies the sound source recognition algorithm based on the L-shaped pickup array structure in the military background. The completed research work is as follows:

(1) The structure of the L-shaped pickup array is analyzed, and the signal preprocessing algorithm of the array node is studied. Based on the LMS adaptive noise cancellation technology, the variable step LMS algorithm that can change the sound spectrum structure of the signal is discussed. It was verified by simulation.
(2) Identify the target sound source to determine the target type. The high-frequency suppression MFCC feature extraction algorithm that can improve the sound spectrum structure of the sound source and the sound source recognition algorithm based on GMM-UBM are studied, and the simulation analysis is carried out. It is proved that the MFCC feature extraction method with high frequency suppression can better distinguish the target sound source, and the GMM-UBM recognition algorithm has a lower score for the non-target sound source category, and its false alarm rate is lower, that is, this algorithm There are fewer nodes that are incorrectly associated.

Acknowledgments. This work was supported by the Natural Science Foundation of Heilongjiang Province [LH2019F017].

References

1. Qin, Y.: Research on Sound Source Localization Technology Based on Microphone Array. Beijing University of Posts and Telecommunications, pp. 1–2 (2019)
2. Deng, Y., Li, J., Zhang, F., Luo, D., Zhu, C., Feng, Z.: Research on improved MUSIC algorithm based on far-field sound source localization. Appl. Electron. Technol. **44**(12), 69–72 (2018)
3. Li, H., Zhou, Y., Liu, H.: Microphone array noise elimination method using phase time-frequency masking. Sig. Process. **34**(12), 1490–1498 (2018)

4. Laufer-Goldshtein, B., Talmon, R., Gannot, S.: Semi-supervised source localization on multiple manifolds with distributed microphones. IEEE Trans. Audio Speech Lang. Process. **25**(3), 1477–1491 (2017)
5. Zhao, C.: Research on adaptive noise canceller. signal processing expert com-mittee of China high-tech industrialization research association. In: Proceedings of the Sixth National Conference on Signal and Intelligent Information Processing and Application. Sig-nal Processing of China High-tech Industrialization Research Association Expert Committee: China High-Tech Industrialization Research Association, pp. 366–368 (2012)
6. Sun, H., Teutsch, H., Mabande, E., et al.: Robust localization of multiple sources in rever-berant environments using EB-ESPRIT with spherical microphone arrays. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic, pp. 117–120 (2011)
7. Li, Y.: Discussion on mel cepstrum MFCC algorithm in speech signal feature extraction. J. Adv. Correspondence Educ. (Nat. Sci. Ed.), **25**(04), 78–80 (2012)
8. Shao, H.J., Zhang, X.P., Wang, Z.: Novel closed-form auxiliary variables based algorithms for sensor node localization using AOA. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1414–1418 (2014)
9. Zong, J., Cui, X.X., Yang, H., et al.: Algorithm and accuracy analysis of weighted maximum likelihood estimation in multi-station DF crossing localization. In: 4th International Conference on Computer, Mechatronics, Control and Electronic Engineering. Atlantis Press (2015)
10. Jingfan, Q., Jingzheng, O.: A novel variable step size LMS adaptive filtering algorithm based on sigmoid function. J. Data Acquisition Process. vol. 3 (1997)