Revija za Kriminalistiko in Kriminologijo
REVIJA ZA KRIMINALISTIKO IN KRIMINOLOGIJO,
ISSN :0034-690X,
STEFANOVA ULICA 2, LJUBLJANA,
SLOVENIA, 1501
Vol. 77 No. 1 (2026)
OPEN ACCESS

Adaptive feature fusion-based speaker recognition strategy for disguised speech speakers

Maolin Ma[1*], Hongbing Zhang[1], Huimin Sun[1]

1. School of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang, Liaoning, 110854, China

aEmail: qd_ma0426@163.com

bEmail: 84707005@qq.com

cEmail: 18262369998@163.com

Funding

This work was supported by Basic scientific research project of colleges and universities, Liaoning Provincial Department of Education: Research on feature pattern extraction and detection method of forged and mutated speech, item number: JYTZD2023150; and 2024 Graduate Innovation Ability Improvement Project, item number: 2024YCYB34.

Abstract: Automatic speaker authentication systems face the threat of disguised speech attacks, especially those generated by speech conversion and speech synthesis techniques, which pose a greater risk to the system and require the development of efficient recognition strategies. This study proposes an adaptive feature fusion-based speaker recognition strategy for disguised speech, which improves the system's ability to detect disguised speech by combining resonance peaks and GFCC feature parameters. The method adopts the inverse spectral method to extract the resonance peak coefficients and combines the GFCC parameter extraction technique, fuses the two features by adaptive weighting, and finally uses a Gaussian mixture model to classify the authentic and fake speech. The experimental results show that in the evaluation set, the average t-DCF of the proposed fusion feature method is only 0.058, which is significantly better than that of the method using the resonance peak feature (0.131)

and the GFCC feature (0.086) alone; in the white noise environment (SNR=20dB), the average equal error rate of the fusion feature method is 11.01%, which is 8.66% lower than that of using the resonance peak feature and GFCC feature alone features by 8.66% and 5.57%, respectively. It is shown that the proposed adaptive feature fusion strategy can effectively improve the performance of the camouflaged speech speaker recognition system, especially in noisy environments, which exhibits stronger robustness.
Keywords: Camouflaged speech; Speaker recognition; Adaptive feature fusion; Resonance peak; GFCC; Anti-spoofing detection

1. Introduction

With the rapid development of computer technology in various fields, artificial intelligence technology has a great role in promoting the development of traditional industries. The research on biometric recognition is the key exploration direction of artificial intelligence nowadays, which has not only achieved rapid development in theoretical research, but also been widely used in practical scenarios [1-2]. Among them, voiceprint recognition, also known as speaker recognition, is a widely accepted and applied recognition technology [3]. With the advantages of easy feature acquisition, good interactivity, and remote authentication, voiceprint recognition has a wide range of applications in the field of information security such as identity authentication systems, personalized applications, access control systems, and security [4-6].

In recent years, the continuous development of speech synthesis technology and voice conversion technology, while enriching people's spare time life, has also provided new ways for criminals to carry out illegal and criminal activities [7-8]. Some of the computer-processed camouflaged speech and the target's real speech are highly similar, even to the extent that it is difficult for the human ear to distinguish, and these camouflaged speech that deliberately imitates the target person poses a great threat to the safe and reliable performance of the voiceprint recognition system, which is one of the challenges facing the promotion and application of voiceprint recognition

technology [9-12]. At present, there has been some progress in the research on the detection of disguised speech. However, most of the researches have only studied and detected one type of spoofed speech, and often the detection rate is not high when dealing with other types of spoofing attacks [13-14]. Therefore, in order to adequately cope with the task of detecting multiple types of disguised speech in complex scenarios, a disguised speech speaker recognition method with adaptive feature fusion is proposed to improve the detection of disguised speech by automatically extracting speech feature information for fusion analysis [15-18].

The rapid development of speech disguise technology poses a serious threat to the automatic speaker authentication system, especially the disguised speech generated by speech conversion and speech synthesis technology is closer to the target speaker in terms of spectral features, which poses a greater hazard to the authentication system. Existing studies have shown that extracting feature parameters with stronger distinguishing ability is the key to improve the performance of the disguised speech recognition system. Based on this problem, researchers have proposed a variety of feature parameter extraction methods, such as amplitude spectrum features and phase spectrum features. However, a single feature often fails to fully capture the characteristics of the disguised speech, especially in noisy environments where the recognition performance is significantly degraded. Considering that the resonance peak as an important parameter describing the acoustic channel can effectively reflect the spectral envelope characteristics of speech, and the GFCC parameter based on the Gammatone filter simulating the cochlear hearing characteristics of the human ear can better suppress the noise interference, this study proposes a strategy of adaptive fusion of these two features. Through the weighted combination of different weighting coefficients, both the human vocal mechanism and the perceptual characteristics of the human ear can be reflected, so as to capture the personality characteristics of the speaker more comprehensively. The study employs the ASVspoof 2019 dataset for validation and introduces a Gaussian mixture model as a classifier, and focuses on exploring the recognition performance of the proposed method under different types of camouflaged speech and noisy environments using t-DCF and equal error rate as evaluation metrics.

2. Speech disguise and recognition fundamentals

2.1 Phony speech

Spoofed voice is usually generated by human imitation, device playback, voice conversion and voice synthesis technology, through these deliberate operations can be disguised as a specific speaker's voice, so as to achieve the purpose of deception of the automatic speaker authentication system, is the current potential threat to the automatic speaker authentication system.

(1) Artificial imitation: Imitation is one of the most obvious ways of deception, which refers to the attack on the automatic speaker authentication system by artificially changing one's own voice, i.e. imitating the voice of the target speaker. The attacker attempts to imitate the timbre and rhythm of the target speaker's voice to deceive the automatic speaker authentication system in the absence of assistive technologies such as computers.

(2) Device playback: Attackers use audio devices to record and collect voice samples from real target speakers and use them to spoof the automatic speaker authentication system.

(3) Speech Transformation: Usually, speech transformation techniques are performed in three stages. First, the input speech is analyzed and features are extracted, then it is converted to match the features of the target speaker's voice, and finally the converted features are re-synthesized into speech using a vocoder.

(4) Speech Synthesis: Speech synthesis, also known as text-to-speech (TTS), converts any input text into corresponding speech. A typical speech synthesis system has two main components: text analysis and speech waveform generation, which are sometimes referred to as front-end and back-end. In the text analysis component, the input text is converted into a linguistic specification consisting of acoustic elements. And in the speech waveform generation component, speech waveforms are generated based on the obtained linguistic specifications.

The spoofed speech generated by the above four techniques poses a greater or

lesser hazard to the current automatic speaker authentication systems. Among them, speech conversion and speech synthesis techniques utilize expertise in speech signal processing, and the disguised speech generated by them is closer to the target speaker's speech in terms of spectral characteristics, posing a greater threat to automatic speaker authentication systems. Therefore, the research on speaker recognition methods for disguised speech carried out in this paper mainly focuses on the disguised speech generated by speech conversion and speech synthesis techniques.

2.2 Recognition fundamentals

Masquerade speech speaker recognition can usually be divided into two phases: the first is the training phase and the second is the detection phase. The block diagram of the camouflaged speech recognition system is shown in Figure 1. In the training phase, feature extraction is performed on all the speech in the real speech library and the camouflaged speech library respectively, and the acquired features are utilized to train classifiers that can distinguish between the real and the camouflaged speech. In the recognition phase, the features of the speech to be detected need to be extracted in advance, and then the trained classifier is applied to recognize the authenticity.

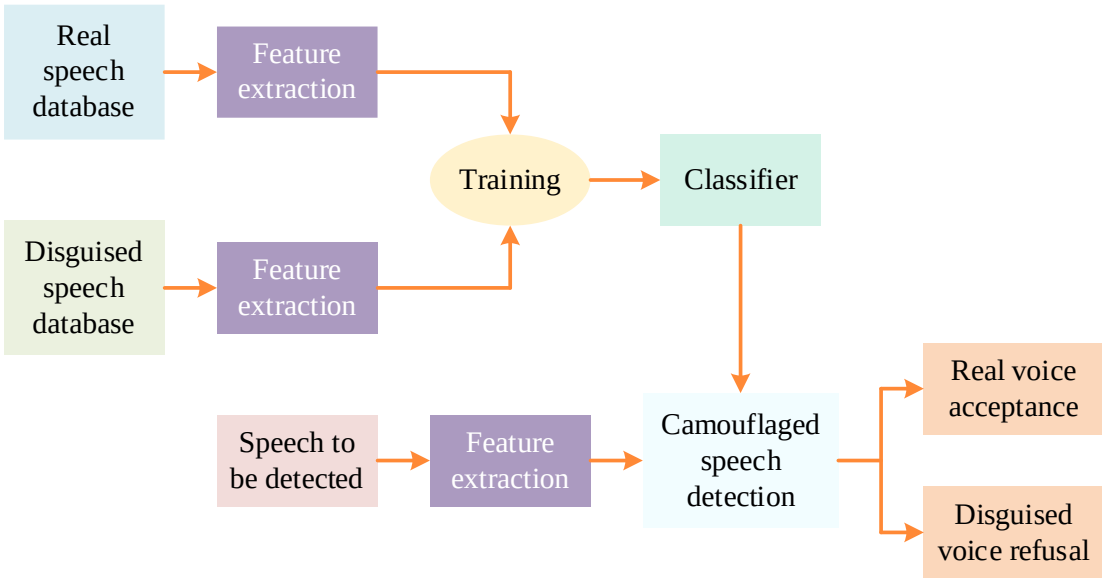


Figure 1 Block of the camouflaged speech detection system

The extracted speech features need to be consistent in the training and detection phases. The features commonly used for artifact recognition are mainly classified into two types, one is the amplitude spectral features of the speech signal, which usually have higher-order Meier inverted spectral coefficients, Meier dominant frequency, and logarithmic amplitude spectra, etc. The other is the phase spectral features of the speech signal. The other is the phase spectrum of the speech signal, usually with modified group delay and relative phase shift. The aim of artifactual speech recognition is to achieve classification of real and artifactual speech, which is essentially a binary classification problem. Therefore training an effective classifier using the obtained features is the key to the camouflage speech recognition system. Commonly used classifiers for disguised speech recognition are Gaussian mixture model, support vector machine and deep neural network. Among them, Gaussian mixture model is easy to implement and shows good performance in disguise speech recognition, so it has been used in many disguise speech recognition systems.

The detection process of the fake speech detection system based on Gaussian mixture model: firstly, the Gaussian mixture model is pre-trained using the real speech library and the fake speech library to build the real speaker speech model λ_{human} and the fake speech model λ_{spoof}, respectively. Then the feature vector O of the speech to be detected is extracted, and then the a posteriori probabilities $p(O|\lambda_{human})$ and $p(O|\lambda_{spoof})$ are computed by applying the real-speaker speech model and the camouflaged speech model in the Gaussian mixture model, respectively. Finally, the log-likelihood ratio $\Lambda(O)$ of the two is calculated and compared with the pre-set threshold θ of the Gaussian mixture model to determine the authentic speech. The mathematical calculation of the log-likelihood ratio $\Lambda(O)$ can be expressed as:

$$\Lambda(O) = \log p(O|\lambda_{human}) - \log p(O|\lambda_{spoof}) \tag{1}$$

where O denotes the feature vector extracted from the speech to be detected, λ_{human} and λ_{spooj} denote the real speaker speech model and the camouflaged speech

model, respectively, and $p(O|\lambda_{human})$ and $p(O|\lambda_{spoof})$ denote the posterior probability of feature vector O under the real speaker speech model and the camouflaged speech model, respectively, and $\Lambda(O)$ denotes the log-likelihood ratio. When the log-likelihood ratio is greater than a threshold θ, it is determined to be real speech, and vice versa for disguised speech.

3. Adaptive feature fusion based camouflage speech recognition strategy

The technical difficulty for voiceprint recognition of disguised speaker's voice is that when the voice is disguised, it can make a huge change in some feature parameters of the voice. Reducing the influence of disguised speech by extracting feature parameters with more distinguishing ability is a key issue to improve the performance of speaker recognition system. Therefore, this paper addresses the problem of low performance of voice recognition system for disguised speech from the perspective of feature extraction. A speaker recognition strategy for disguised speech based on adaptive feature fusion is proposed using the fusion of resonance peaks and GFCC feature parameters.

3.1 Extraction of resonance peak coefficients by inverse spectroscopy

3.1.1 Resonance peaks

Resonance peak is one of the important feature parameters of voiceprint recognition. The parameters of the resonance peak include the resonance peak frequency and bandwidth, the spectral envelope of the vocal tract information is roughly the same as the spectral envelope of the speech information, so extracting the resonance peak is to get the spectral envelope of the speech and take the great value in the spectral

envelope as the resonance peak parameter. In this paper, the resonance peaks of speech are extracted based on the inverse spectral method. Firstly, the homomorphic analysis method is used to eliminate the influence of excitation, obtain the information of the vocal tract part, and finally find the resonance peak of speech.

3.1.2 Resonance peak extraction

When the signal sequence is $x(n)$, its Fourier transform is:

$$X(w) = FT[x(n)] \tag{2}$$

Then the sequence:

$$\hat{x}(n) = FT^{-1}[\ln | X(w) |] \tag{3}$$

Call $\hat{x}(n)$ the cepstrum, i.e., the sequence of cepstrums of $x(n)$. $\hat{x}(n)$ is the Fourier inverse transform of $x(n)$. where FT and FT^{-1} denote the Fourier transform and Fourier inverse transform, respectively.

The homomorphic deconvolution technique is used to separate the fundamental tone information from the vocal tract information in the inverse spectral domain, and the resonance peaks of the speech are extracted based on the vocal tract information. The speech $x(n)$ is obtained from the glottal pulse $e(n)$ filtered by the acoustic channel response $h(n)$ as shown in equation (4):

$$x(n) = e(n) * h(n) \tag{4}$$

The inverse spectrum is solved for the speech signal to obtain equation (5):

$$\hat{x}(n) = \hat{e}(n) + \hat{h}(n) \tag{5}$$

It can be concluded that the fundamental information $\hat{e}(n)$ and the vocal tract information $\hat{h}(n)$ in the cepstrum domain can be considered to be relatively independent. With the cepstrum method, $e(n)$ and $h(n)$ can be separated, and then

the resonance peaks can be obtained according to the excitation $h(n)$ and the characteristics of the cepstrum. The specific steps are as follows:

(1) Pre-emphasize the sound signal $x(n)$, and after adding windowed sub-frames (with frame length N), obtain $x_i(n)$, where i denotes the i th frame of the sound signal.

(2) The discrete Fourier transform of $x_i(n)$ is obtained:

$$X_i(k) = \sum_{n-0}^{N-1} x_i(n) e^{-j2\pi kn/N} \tag{6}$$

(3) Taking the amplitude for $X_i(k)$ and then taking the logarithm gives:

$$\hat{X}_i(k) = \log(|X_i(k)|) \tag{7}$$

(4) Perform a Fourier inverse transform on $\hat{X}_i(k)$, from which the cepstrum sequence is obtained:

$$\hat{x}_i(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}_i(k) e^{j2\pi kn/N} \tag{8}$$

(5) Set a low-pass window function $window(n)$ on the cepstrum domain axis, which can generally be set as a rectangular window:

$$window(n) = \begin{cases} 1, n \le n_0 - 1 and n \ge N - n_0 + 1 \\ 0, n_0 - 1 < n < N - n_0 + 1 \end{cases} \tag{9}$$

where n_0 is the width of the window function. The window function is then multiplied by the inverse spectral sequence $\hat{x}(n)$ to obtain:

$$h_i(n) = \hat{x}_i(n) \times window(n) \tag{10}$$

(6) Putting $h_i(n)$ through the Fourier transform gives the envelope of $X_i(k)$:

$$H_i(k) = \sum_{n=0}^{N-1} h_i(n) e^{-j2\pi kn/N} \tag{11}$$

(7) Find the maximum value on the envelope to get the resonance peak parameters.

3.2 Extraction of GFCC parameters

3.2.1 Gammatone filters

The Gammatone filter simulates the cochlear hearing model of the human ear and can well represent the crossover characteristics of the basilar membrane, while effectively suppressing the interference of noise. Its time domain expression is as follows:

$$h(t) = kt^{n-1}e^{-2\pi bt}\cos(2\pi f_c t + \phi), t \geq 0 \tag{12}$$

where ϕ is the phase, f_c is the center frequency, n denotes the order of the filter, when $n = 3,4,5$, the Gammatone filter can better simulate the auditory characteristics of the basilar membrane of the human ear, k is the filter gain, b is the attenuation factor, which is related to the bandwidth of the filter, and which controls the decay rate of the impulse response, and the center The relationship between f and the center frequency is:

$$b = 1.019 * 24.7 * (4.37 * f_c / 1000 + 1) \tag{13}$$

Equation (13) consists of two parts: the filter envelope $kt^{n-1}e^{-2\pi bt}$ and the modulation of the frequency f_c by $\cos(2\pi f_c + \phi)$. It can be thought of as the product of a Gamma distribution and a cosine signal, or as a cosine signal modulating the Gamma distribution to the frequency f_c.

By Fourier transform, the frequency domain expression for $h(t)$ is given by the following equation:

$$H(f) = \frac{k(n-1)!}{2(2\pi b)^n}\left\{\left(\frac{j(f-f_c)}{b}+1\right)^{-n} + \left(\frac{j(f+f_c)}{b}+1\right)^{-n}\right\} \tag{14}$$

When f_c / b is large enough, $\left[j(f+f_c)/b+1\right]^{-n}$ can be ignored.

Let $s = j2\pi f$, then its Laplace transform is expressed as:

$$H(s)=\frac{k(n-1)!}{2}(s-(j2\pi f_c-2\pi b))^{-n} \tag{15}$$

Transform it by Z as:

$$H(z)=\frac{k(n-1)!}{2}(1-e^{j2\pi f_c-2\pi b}z^{-1})^{-n} \tag{16}$$

Order for:

$$A(z)=\frac{1}{1-e^{j2\pi f_c/f_s-2\pi b/f_s}z^{-1}} \tag{17}$$

$H(z)$ can be thought of as a cascade of recursive applications of $A(z)$. It can be simplified by cascading the application of a series of filters that first remove the f_c component, then pass it through a base filter $\hat{H}(z)$ independent of f_c, and in the end compensate for f_c.

3.2.2 GFCC Feature Extraction

After the speech signal is preprocessed and passed through a Gammatone filter bank based on the cochlear hearing properties of the human ear, a set of cepstrum feature parameters can be obtained, which is notated as GFCC, and in turn can be used in a speaker recognition system. The overall steps of feature extraction are:

(1) Pre-processing.

(2) Fast Fourier Transform: Fast Fourier Transform is performed on the preprocessed speech signal to transform the time domain signal into frequency domain signal to obtain the power spectrum.

(3) Filtering: the energy spectrum obtained by squaring the power spectrum is filtered by a Gammatone filter bank.

(4) Logarithmic compression: Logarithmic compression is applied to the output of each filter to further simulate the nonlinear characteristics of human speech perception.

(5) Discrete cosine transformation: The energy spectrum after logarithmic compression is discretized by cosine transformation to uncorrelate it and obtain better

energy compression.

The first-order difference and second-order difference are selected as the dynamic features, and the GFCC is combined with the first-order difference and second-order difference to get the feature vector about the GFCC.

3.3 Adaptive feature fusion

Since the resonance peak is one of the important parameters for describing the vocal tract in speech signal processing, meanwhile the GFCC embodies a kind of auditory property that simulates the human ear. Therefore, the combination of the two types of speech feature parameters weighted with different weighting coefficients can reflect both the mechanism of human vocalization and the perceptual characteristics of the human ear, and the combination of human vocalization and hearing, as the basis for the modeling of acoustic models, can better reflect the personality characteristics of the speaker, and the process of adaptive feature parameter fusion is shown in Fig. 2.

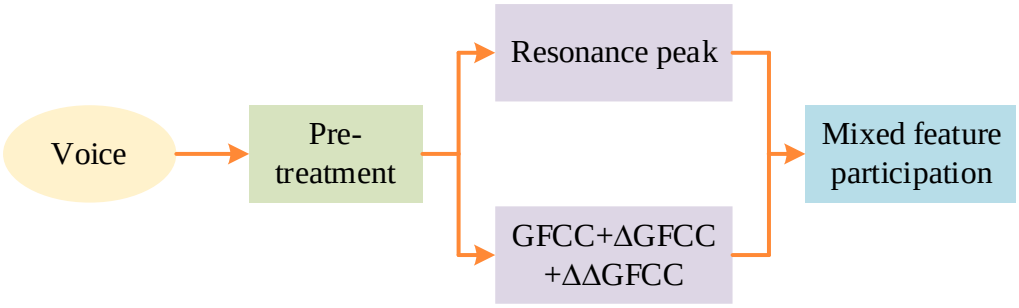


Figure 2 Block of combined feature parameters

Firstly, the speech signal is preprocessed to extract the vocal tract impulse information $\hat{h}(n)$ and G_m (GFCC as well as the first-order difference and the second-order difference of the GFCC) obtained by Gammatone filter in speech, respectively, and after that, these two kinds of feature parameters are combined to form a mixture of features as inputs to the acoustic model. After that, $\hat{h}(n)$ and G_m are to be normalized as shown in Eqs. (18) and (19):

$$\hat{h}(n)^{'} = \frac{\hat{h}(n)}{\hat{h}(n)_{\max}} \tag{18}$$

$$G_m^{'} = \frac{G_m}{G_{m-\max}} \tag{19}$$

Where $\hat{h}(n)_{\max}$ is the maximum value of the resonance peak eigenparameter obtained from the test, and $G_{m-\max}$ is the maximum value of the GFCC and its differential eigenparameter obtained from the test. After this treatment, $\hat{h}(n)^{'}$ and $G_m^{'}$ are both between 0 and 1, and then let:

$$d_1 = \hat{h}(n)^{'} \tag{20}$$

$$d_2 = G_m^{'} \tag{21}$$

The impact factors of the two methods in Eqs. (22) and (23) can be expressed using the average of the test set:

$$C_1 = \frac{ave(d_1)}{(ave(d_1) + ave(d_2))} \tag{22}$$

$$C_2 = \frac{ave(d_2)}{(ave(d_1) + ave(d_2))} \tag{23}$$

Where C_1 is the coefficient factor of the resonance peak, and C_2 is the coefficient factor of the GFCC and its differential features, which represent the influence of the two kinds of feature parameters on the recognition results, respectively. The final hybrid feature parameter obtained is the weighted combination of the two feature parameters, as shown in equation (24):

$$S = C_1\hat{h}(n)^{'} + C_2 G_m^{'} \tag{24}$$

3.4 Gaussian mixture models

Since the study of disguised speech relies on speaker recognition technology, and Gaussian Mixture Model (GMM) can effectively characterize the vocal differences of different speech, in order to make full use of the characteristics of the mixture of feature

parameters to identify the disguised speech, so based on the GMM model of the fusion of the obtained feature parameters to be modeled.

Gaussian mixture model can be used to classify the fake speech and real speech, Gaussian mixture model is often used in the fake speech detection system to classify the fake speech and real speech, it is a kind of model commonly used in probability statistics.

By the central limit law and the large number theorem, the linear weighting of M Gaussian probability density distribution functions can model any continuous probability distribution. For a d-dimensional feature vector X, the kth Gaussian component satisfies the normal distribution $X \sim N(m_k, \Sigma_k)$, then the form of the GMM can be defined as follows:

$$p(X \mid \lambda) = \sum_{k=1}^{M} \pi_k N(X \mid m_k, \sum_k) \tag{25}$$

The weight coefficients attributed to each Gaussian component in the GMM model are k, where $k = 1, 2, ..., M, \pi_k, \sum_{k=1}^{M} \pi_k = 1$. The average vector of individual Gaussian weights is denoted by m_k, and then its covariance matrix Σ_k can be obtained by computing the Gaussian weights, and λ denotes the set of parameters of the Gaussian mixture model, i.e., the set of parameters of the Gaussian mixture model can be obtained by means of $\lambda = \{\pi_k, m_k, \Sigma_k\}$ parameters a GMM model can be represented. The $N(\cdot)$ denotes the Gaussian density function, which can be expressed mathematically as:

$$N(X \mid m_k, \sum_k) = (2\pi)^{-\frac{D}{2}} \mid \sum_k \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - m_k)^T \sum_k^{-1} (x - m_k)\right\} \tag{26}$$

In terms of computational effort, diagonal matrices are more suitable for inverse computations when the dimensionality is high, so Σ_k is chosen for such matrices, i.e:

$$\sum_k = diag\left\{\sigma_{k_0}^2, \sigma_{k_1}^2, ..., \sigma_{k_{D-1}}^2\right\} \tag{27}$$

where $\sigma_{k_i}^2$ denotes the variance of the d th-dimensional component of the GMM model when the corresponding eigenvector of the subdistribution is k, where $d = 0,1,\ldots,D-1$. Then the Gaussian probability density function $f_k(X)$ is:

$$f_k(X) = \prod_{d=0}^{D-1} \frac{1}{\sqrt{2\pi}\delta_{k_d}} \exp\left\{ -\frac{(X_d - m_{k_d})^2}{2\delta_{k_d}^2} \right\} \tag{28}$$

where X_d is the d th component of x and m_{k_d} is the d th component of m_k.

The GMM training process uses the Expectation Maximization Algorithm (EM).The main idea of EM is iterative, constantly predicting the unobtained situation by taking the acquired situation and then launching a new model with the acquired situation.

3.5 Performance evaluation indicators

3.5.1 Equal error rate

The equal error rate is usually used for the pseudo-speech detection system and is an effective performance evaluation criterion, and the detection capability of the system can be seen more directly by comparing the equal error rate.

(1) Error acceptance rate

In the disguise speech detection system processing classification results, the error acceptance rate P_{FAR} indicates the ratio of the number of samples of the system in verifying the number of samples of the deceptive speech incorrectly accepted speech to the number of samples of the disguise speech, which is expressed by the formula as:

$$P_{FAR} = \frac{\text{The number of wrongly accepted disguised voice samples}}{\text{The number of disguised voice samples}} \times 100\% \tag{29}$$

(2) False rejection rate

When the spoofed speech detection system processes the classification results, the false rejection rate P_{FRR} indicates that the system incorrectly accepts the spoofed speech in verifying the

The ratio of the number of samples of speech to the number of samples of spoofed speech is expressed by the formula:

$$P_{FRR}=\frac{\text{The number of real voice samples that were wrongly rejected}}{\text{The number of real speech samples}}\times 100\% \quad (30)$$

The error acceptance rate decreases monotonically as the threshold increases, while the error rejection rate increases monotonically as the threshold increases. Due to this property of theirs, they are bound to coincide when a certain threshold is obtained, in which case the equal error rate EER is obtained, at which point the magnitude of the value of the false rejection rate is the same as the magnitude of the value of the false acceptance rate, and hence the equal error rate can be mathematically expressed as:

$$EER = P_{FAR} = P_{FRR} \quad (31)$$

The false acceptance rate and false rejection rate should be as small as possible in the evaluation phase of a speech detection system, so the smaller the value of the equal error rate can indicate that the system is more capable of detection. The equal error rate is also often used as an evaluation criterion for spoofed speech detection systems.

3.5.2 Tandem Detection Cost Functions

In the task of anti-spoofing attack, when the anti-spoofing attack system is cascaded with the automatic speaker verification (ASV) system, the EER often cannot be used as a reliable performance evaluation criterion, if the EER of the binary classification error is considered alone is insufficient to measure the overall anti-spoofing attack system and the ASV system, because if the real speech is judged to be true by the early judgment error, then the ASV system will possibly lose a sentence of the target speech. Similarly, if the spoofed speech is judged true by the anti-spoofing attack system, then the ASV system will face the attack of spoofed speech, which will

also lead to a decrease in the EER of the ASV system. Therefore, the tandem detection cost function (t-DCF) is used, which integrates the binary classification decision and speaker recognition, and also considers the relationship between the anti-spoofing attack system and the ASV in tandem order or the CM and the ASV in parallel: a true speech will exist as a target or non-target, and a spoofed speech in the same way. The formula is as follows:

$$t-DCF = C_{FRR}\pi_{tar}P_{FRR}^{CM} + C_{FAR}(1-\pi_{tar})P_{FAR}^{CM} \tag{32}$$

In the formula, the a priori probability π_{tar} =0. 9405, ASV surrogate value $C_{FRR}=1, C_{FAR}=10$. In detection, the smaller t-DCF value represents the better detection effect of the whole artifact speech detection system.

4. Experiments and analysis of results

4.1 Data sets

The database provided by the ASVspoof 2019 challenge, which is based on the VCTK corpus and created from the speech of 107 speakers, was selected for the experiments. In the ASVspoof 2019 database, the database is divided into Logical Access (LA) scenario and Physical Access (PA) scenario according to the use case scenarios, and three kinds of spoofed speech, namely TTS, VC, and Replay Attack, are considered comprehensively.In the LA scenario, spoofing attack is a system-level attack, which occurs during the system's processing of speech signals up to the decision-making process, where spoofed speech is generated through the state-of-the-art TTS and VC techniques.In the PA scenario, the spoofing attack is a microphone-level attack, and the spoofed speech is directly used as an input to the ASV system.The spoofed speech in this scenario consists of well-prepared, high-quality replayed speech. The focus of this paper is to propose an efficient recognition method for the disguised speech speaker, so the LA dataset is selected for all the experiments.

The LA dataset includes attacker data and defense data, and this paper adopts the

defense data to carry out the research. Among them, the training set (Train) includes 2580 real speech and 22800 disguised speech, involving 4 TTS algorithms (A01~A04) and 2 VC algorithms (A05~A06). The development set (Dev) is consistent with the algorithms used in the training set and includes 2548 real speech and 22,296 disguised speech. The evaluation set (Eval) contains 7355 real speech and 63882 disguised speech, involving 10 TTS algorithms (A07~A16). Since the aim of this paper is to study the detection methods for synthetic speech, the experiments all use the relevant parts of the TTS algorithms in each subset.

4.2 Experimental results

4.2.1 t-DCF analysis

The use of resonance peak (RP) feature vectors, GFCC feature vectors and this paper's fusion of RP-GFCC feature vectors can be realized on the synthesized speech camouflage detection, will be these methods for the development of the development set and evaluation of various types of camouflage speech speaker recognition, development of different types of camouflage speech in the development of the set of the detection of the t-DCF comparisons are shown in Fig. 3, the evaluation of the set of different types of camouflage speech detection of the t-DCF comparison is shown in Fig. 4. The t-DCF comparison of the detection of different types of disguised speech in the development set is shown in Figure 4. For different types of disguised speech, the recognition method that incorporates resonance peaks and GFCC feature parameters in this paper has the best detection effect. In both the development and evaluation sets, the t-DCF results of this paper's method for the 14 camouflaged speech are lower than those of the other two feature extraction methods, with the mean values of t-DCF results being 0.062 and 0.058, while the mean values of t-DCF results for the resonance peak feature vectors are 0.119 and 0.131, and those for the GFCC feature vectors are 0.094 and 0.086. It shows that this paper's adaptive feature fusion-based fake speech recognition strategy further improves the detection performance of fake speech speakers.

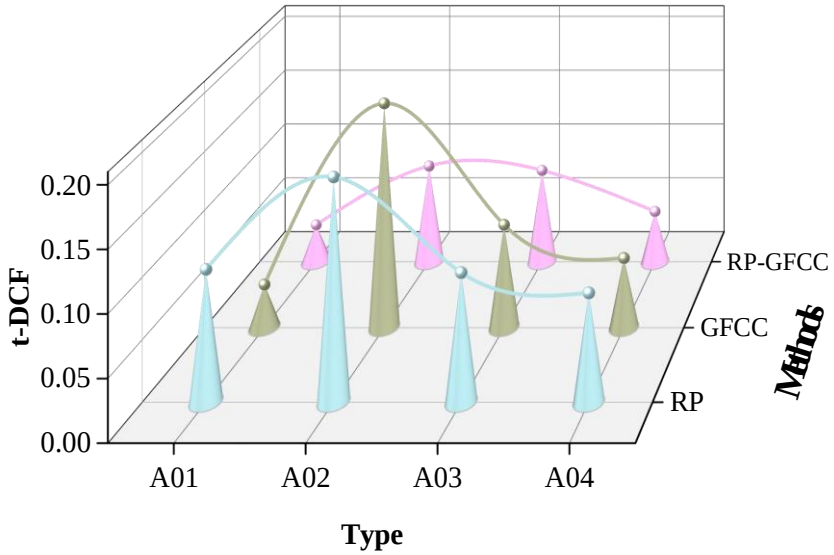


Figure 3 Comparison of t-DCF in different types of camouflage speech (Development set)

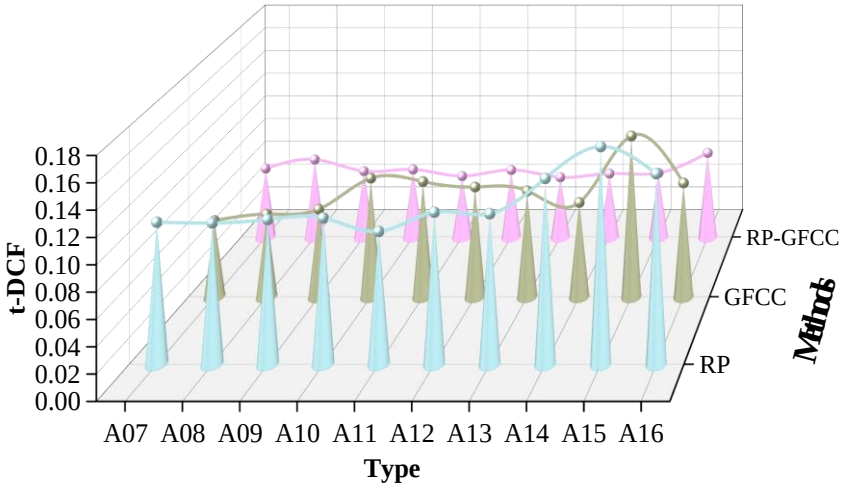


Figure 4 Comparison of t-DCF in different types of camouflage speech (Evaluation set)

4.2.2 Equal error rate analysis

This section explores the performance of three feature extraction methods for camouflaged speech speaker recognition under WHITE noise. The noise corpus used in the experiment is obtained by summing the pure speech and noise speech in the ASVspoof2019 corpus, and the noise speech is selected from the noisex-92 noise library, of which white noise is selected in this paper, and the signal-to-noise ratio (SNR) is selected from the three types of 0dB, 10dB and 20dB. The equal error rate is used as the evaluation index of the spoofed speech detection system. Table 1 shows the equal error rate of each camouflaged speech in the development set under white noise, and

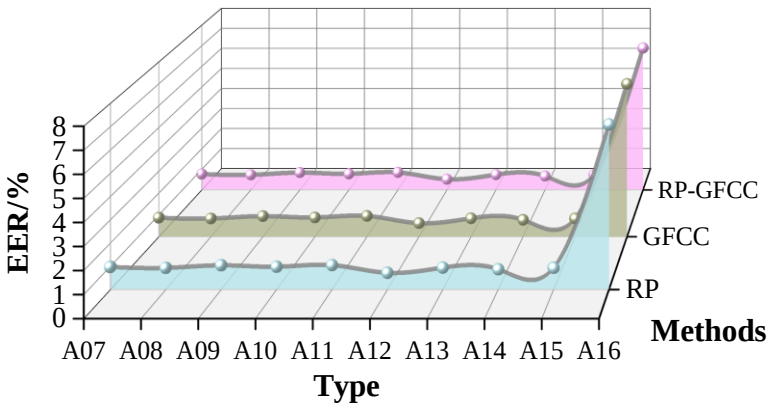
Fig. 5 shows the equal error rate of each camouflaged speech in the evaluation set under white noise.

The EER of the camouflaged speech speaker recognition system is significantly increased after the addition of white noise, indicating that the performance of the system is greatly degraded. When SNR=0, the performance of the recognition model trained based on RP-GFCC fusion features is less different from the performance of the model trained based on RP features and based on GFCC features, and the recognition is yet to be improved, with a difference in EER of 3.17% and 1.58%. When the SNR is increased, the EER of the recognition model trained on RP-GFCC fusion features is lower than that of the model trained on RP features and GFCC-based features, indicating that the algorithm outperforms other feature extraction methods in recognizing disguised speech speakers. When the SNR is 10 and 20, the EER of the feature fusion recognition method in this paper is 5.66% and 3.38%, 8.66% and 5.57% lower than the models trained on RP features and GFCC based features. From the figure, it can be seen that the RP-GFCC fusion feature based camouflage speech recognition model is better than the RP and GFCC based camouflage speech recognition model in detecting under noisy conditions in all cases, and the equal error rate of the three methods increases significantly when facing a camouflage speech attack like A16. This is due to the fact that the camouflaged speech like A16 is generated by a synthesis algorithm based on unit selection, which generates the camouflaged speech by selecting speech segments and cascading them to generate the camouflaged speech, which retains more recognizable features due to the fact that the selected speech segments are the speech of the real speaker, resulting in a higher equal error rate of the camouflaged speech speaker recognition model when facing such an attack.

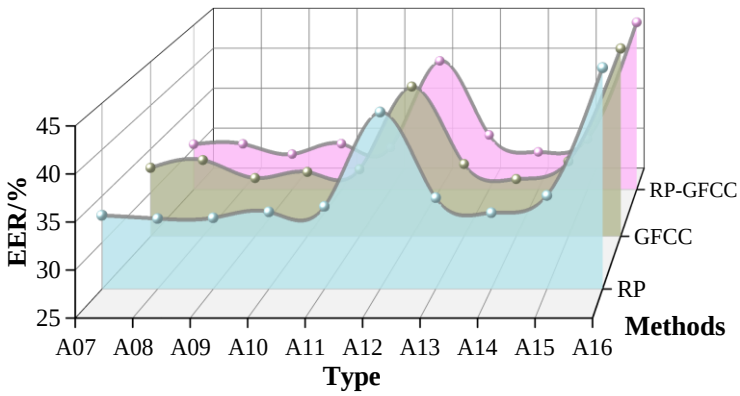
Table 1 EER of each disguised voice under white noise of development set

SNR	Feature extraction	EER(%)				
		A01	A02	A03	A04	Average
Clean	RP	0.67	0.97	0.17	0.30	0.53
	GFCC	0.52	0.61	0.13	0.26	0.38

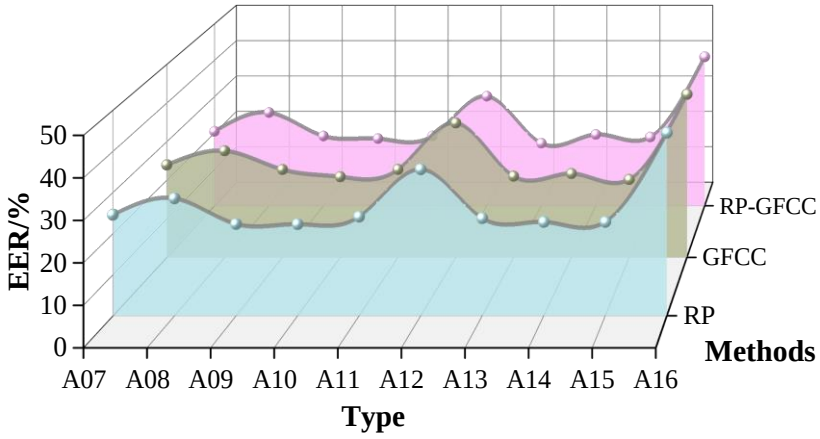
	RP-GFCC	0.48	0.53	0.11	0.18	0.33
SNR=0	RP	30.50	33.97	29.05	31.02	31.14
	GFCC	29.52	31.05	28.36	29.28	29.55
	RP-GFCC	28.40	27.54	27.52	28.42	27.97
SNR=10	RP	28.33	27.54	23.81	21.27	25.24
	GFCC	27.16	25.07	20.47	19.13	22.96
	RP-GFCC	22.47	22.07	18.43	15.33	19.58
SNR=20	RP	20.66	20.97	18.48	18.57	19.67
	GFCC	17.39	19.55	14.63	14.73	16.58
	RP-GFCC	12.78	11.24	10.17	9.83	11.01



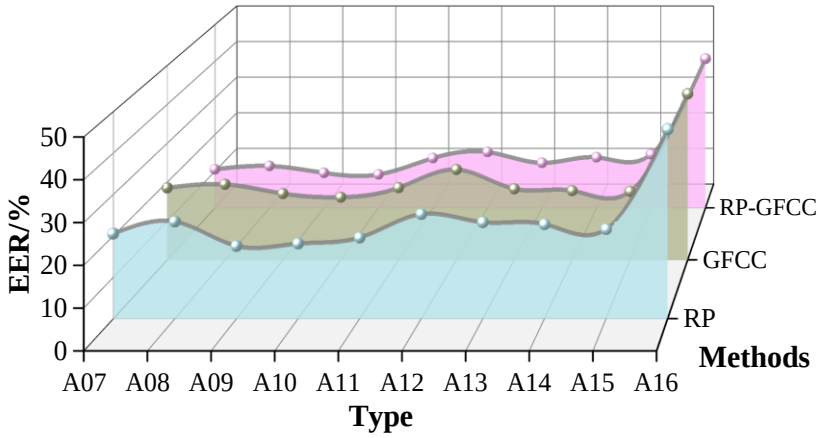
(a)Clean



(b)SNR=0



(c)SNR=10



(d)SNR=20

Figure 5 EER of each disguised voice under white noise of evaluation set

5. Conclusion

In this paper, we proposed an adaptive feature fusion-based strategy for camouflaged speech speaker recognition, and conducted a systematic study for camouflaged speech recognition system by combining resonance peak and GFCC feature parameters. The experimental results show that the average equal error rate of the proposed method on the development set is only 0.33% in a noise-free environment, which is 0.20% and 0.05% lower than the use of resonance peak features and GFCC features alone, respectively. For different types of camouflaged speech, the average value of t-DCF results of this method on the development set for 14 types of camouflaged speech is 0.062, which is significantly lower than the results of the

resonance peak feature (0.119) and the GFCC feature (0.094). In the noise immunity performance test, the average equal error rate of this method is 19.58% when the SNR is 10 dB, which is 5.66% and 3.38% lower than the other two methods, respectively. It is worth noting that for the A16-type camouflaged speech generated based on the unit selection synthesis algorithm, the equal error rates of the three methods are higher, which is due to the fact that this type of camouflaged speech retains more features of the real speaker. The results demonstrate that the adaptive feature fusion strategy not only effectively improves the recognition performance of disguised speech, but also shows strong robustness in various noise environments, which provides important technical support for the security of automatic speaker authentication systems.

About the Authors

Maolin Ma was born in Qingdao, Shandong. P.R. China, in 1998. He obtained a bachelor's degree from Guilin University of Electronic and Technology in China. I am currently studying at the School of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China. My main research direction is Speaker Recognition and Voice Anti-Spoofing.

Hongbing Zhang was born in Wuyang, Henan. P.R. China, in 1979. He obtained a bachelor's degree from Shaanxi Normal University, Xi'an in China. I am currently a Professor at the School of Police Information Technology and Intelligence, Criminal Investigation Police University of China. My main research direction is Speaker Recognition and Voice Anti-Spoofing.

Huimin Sun was born in Yancheng, Jiangsu. P.R. China, in 2000. He obtained a bachelor's degree from Dalian Jiaotong University in China. I am currently studying at the School of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China. My main research direction is speech emotion recognition.

References

[1] Iyer, A. P., Karthikeyan, J., Khan, R. H., & Binu, P. M. (2020). An analysis of artificial intelligence in biometrics-the next level of security. J Crit Rev, 7(1), 571-576.

[2] Abdullahi, S. B., Khunpanuk, C., Bature, Z. A., Chiroma, H., Pakkaranang, N., Abubakar, A. B., & Ibrahim, A. H. (2022). Biometric information recognition using artificial intelligence algorithms: A performance comparison. IEEE Access, 10, 49167-49183.

[3] Li, X., & Mills, M. (2019). Vocal features: from voice identification to speech recognition by machine. Technology and culture, 60(2), S129-S160.

[4] Wang, P., Li, J., Wang, H., Chen, H., Cao, J., Xu, Y., & He, J. (2022, February). Intelligent Access Control System Based on Voiceprint and Voice Technology. In 2022 11th International Conference of Information and Communication Technology (ICTech)) (pp. 461-465). IE

[5] Xavier, L. A. (2019). Identification of Age Voiceprint Using Machine Learning Algorithms. ResearchBerg Review of Science and Technology, 1(1), 1-16.

[6] Ma, Y. (2023). Voiceprint recognition and cloud computing data network security based on scheduling joint optimisation algorithm. International Journal of Global Energy Issues, 45(6), 602-626.

[7] Smith, A. B., Mason, N., Browne, M. E., & Sullivan, B. (2019). Acoustic characteristics of disguised speech. The International Journal of Speech, Language and the Law, 26(1), 85-95.

[8] Staroniewicz, P. (2024). Subjective tests of speaker recognition for selected voice disguise techniques. International Journal of Electronics and Telecommunication, 70(3), 615-620.

[9] Gui, S., Zhou, C., Wang, H., & Gao, T. (2023). Application of Voiceprint Recognition Technology Based on Channel Confrontation Training in the Field of Information Security. Electronics, 12(15), 3309.

[10] Boles, A., & Rad, P. (2017, June). Voice biometrics: Deep learning-based voiceprint authentication system. In 2017 12th system of systems engineering

conference (SoSE) (pp. 1-6). IEEE.

[11] Lou, J., Xu, Z., Zuo, D., Zhang, Z., & Ye, L. (2021). Audio information camouflage detection for social networks. Frontiers in Physics, 9, 715465.

[12] Sharma, R., Govind, D., Mishra, J., Dubey, A. K., Deepak, K. T., & Prasanna, S. R. M. (2024). Milestones in speaker recognition. Artificial Intelligence Review, 57(3), 58.

[13] Lin, Y. S., Chen, H. Y., Huang, M. L., & Hsieh, T. Y. (2024). Data Augmentation for Voiceprint Recognition Using Generative Adversarial Networks. Algorithms, 17(12), 583.

[14] Velayuthapandian, K., & Subramoniam, S. P. (2023). A focus module-based lightweight end-to-end CNN framework for voiceprint recognition. Signal, Image and Video Processing, 17(6), 2817-2825.

[15] Shen, Q., Guo, M., Huang, Y., & Ma, J. (2024). Attentional multi-feature fusion for spoofing-aware speaker verification. International Journal of Speech Technology, 27(2), 377-387.

[16] Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. Expert Systems with Applications, 90, 250-271.

[17] Karthikeyan, V., & Suja Priyadharsini, S. (2024). A stacked convolutional neural network framework with multi-scale attention mechanism for text-independent voiceprint recognition. Pattern Analysis and Applications, 27(2), 48.

[18] Sun, W. Z., Wang, J. S., Zheng, B. W., & Li, Z. F. (2021). A novel convolutional neural network voiceprint recognition method based on improved pooling method and dropout idea. IAENG International Journal of Computer Science, 48(1), 202-212.