Article

Not peer-reviewed version

Application of Speech Recognition Technology Based on Channel Adversarial Training in the Field of Speech Recognition System

Suying Gui , Chuan Zhou * , Hao Wang , Tiegang Gao

Posted Date: 24 July 2023

doi: 10.20944/preprints202307.1591.v1

Keywords: voice recognition; channel adversarial training; information security domain; speaker confirmation

Article

Application of Speech Recognition Technology based on Channel Adversarial Training in the Field of Speech Recognition System

Suying Gui [1,#], Chuan Zhou [2,3,*,#], Hao Wang [4] and Tiegang Gao [1]

[1] College of Software, Nankai University, Tianjin 300100, China
[2] School of Microelectronics, Tianjin University, Tianjin 300100, China
[3] China United Network Communication Group Co., Ltd., Beijing 100000, China
[4] Education Foundation of Beijing Central University for Nationalities, Beijing 100000, China
[*] Correspondence: zhouchuan@tju.edu.cn
[#] The authors contributed equally to this work

Abstract: With the rapid development of big data, artificial intelligence, and Internet technologies, human-human contact and human-machine interaction have produced an explosive growth of voice data. Rapidly identifying the speaker's identity and retrieving and managing his or her speech data in the massive amount of speech data has become a major challenge for intelligent speech applications in the field of information security. This research proposes a vocal recognition technique based on information adversarial training for speaker identity recognition in massive audio and video, and speaker identification when oriented to the information security domain. The experimental results show that the method projects data from different scene channels all onto the same space and dynamically generates interactive speaker representations. It solves the channel mismatch problem and effectively improves the recognition of the speaker's voice patterns across channels and scenes. It is able to separate overlapping voices when multiple people speak at the same time and reduce speaker separation errors. It realizes speaker voice recognition for the information security field and achieves an 89% recall rate in a massive database, which has practical application value for the intelligent application field.

Keywords: voice recognition; channel adversarial training; information security domain; speaker confirmation

1. Introduction

With the development of big data, the Internet, and artificial intelligence technologies, human-computer interaction and communication for all have become globalized, and with the increasing frequency of these communications, the speech data generated shows an explosive growth trend [1]. In the field of information security-oriented and intelligent speech, how to quickly locate the identity of the speaker in the massive speech data, and confirm, retrieve, and manage the speaker's speech has become a major problem in practical applications [2]. For example, the management of blacklisted people in intelligent customer service, voice payment in online banking, and the control, identity confirmation, and attendance authentication of people using voice information in the Internet domain [3]. Human-to-human communication and human-computer interaction are the most direct and convenient ways, which also encompass a large amount of information. The most common content information is one of them, and the most generalized nowadays is voice information [4]. Voice information which includes linguistic information, speaker information, emotional information, and environmental information. Voice recognition, also known as speaker recognition, is based on the speaker identity information contained in the voice, and the speech content information [5]. This research addresses text-independent speaker recognition in the field of information security for massive audio and video, as well as for text-related speaker confirmation scenarios for speaker identity confirmation. The text-irrelevant speaker recognition based on channel adversarial training can effectively solve the scenario channel mismatch between registered speech and test speech, and improve the speech recognition performance. The channel adversarial training is used to extract speaker representations to improve speaker recognition. The attention system is then influenced based on the registered and tested speech to improve the confirmation effect with a dynamic interaction and speaker features. The research is divided into three main parts, the first part is oriented to the overview of the vocal databases in the field of information security, which involves the storage of voice vocal databases and the storage and management of massive speaker data. The composition of the speaker recognition system for information security, the establishment of the speaker voice database, and the framework design of the recognition system. The second part is a text-independent speaker recognition method based on channel adversarial training and the design of the channel adversarial training recognition network structure. The third part is the experimental verification and analysis of the performance of the channel adversarial training-based voice recognition technique to demonstrate the effectiveness of the proposed method in this study.

2. Related Works

With the development of artificial intelligence technology, voice has direct and convenient characteristics, and therefore becomes the preferred way of communication, and vocal recognition technology has a non-negligible role in the huge amount of voice data generated. Sun et al. proposed an improved convolutional neural network-based vocal recognition method. By introducing an improved pooling method, the activation values are squared after the activation function and squared probabilities are assigned to achieve random pooling. The advantages of random pooling are combined while retaining the feature extraction of the maximum pooling method. The possibility of extracting hidden features is enhanced, and the effectiveness of the method is verified by simulation results using a grayscale digital acoustic spectrometer in a self-built acoustic database for acoustic recognition experiments [6]. Hong Z et al. proposed an end-to-end acoustic recognition algorithm based on a convolutional neural network, which uses convolution and downsampling of the convolutional neural network to preprocess the speech signal in end-to-end acoustic recognition. From the preprocessed signal, feature parameters are extracted and a background model is used to model the voiceprint recognition model [7]. Zhang et al. proposed a method that combines differentiable architecture search (Differentiable Architecture Search, DARTS) with generative adversarial learning (Generative Adversarial Nets, GAL). In order to reduce the human and repeated resource consumption required to design the neural network architecture, and make the task to find the best network architecture with better performance more easy to implement. The test accuracy of this method on the voice print recognition dataset is higher than that of DARTS, and the experimental results show that the accuracy on CIFAR10 improves by 7.35%, reaching the state-of-the-art results of 99.60%, with the highest pruning rate of 62.3% [8].

With the popularity of smartphones and other smart devices, personal speech is stored and transmitted on different hardware and applications, and techniques for cross-channel applications have received increasing attention. Yu et al. proposed a technique for race, age, and gender image transformation by generating adversarial networks. The dataset used for age estimation studies suffers from age class imbalance due to the different age distribution of race or gender. This leads to one-sided overfitting of the training data and reduces the generalizability of age estimation. Experiments using four open databases have shown that the method outperforms state-of-the-art methods [9]. Wang et al. proposed a spatial domain and channel domain attention mechanism addressed the problem of manually designed feature extraction and fusion rules in existing image fusion methods. Experimental results show that the detection model trained by transfer using fused images has the best performance with accuracy P, recall R, average accuracy, and F-1 scores of 0.804, 0.923, 0.928, and 0.859, respectively [10]. Zhu et al. proposed a hybrid Acoustic model to solve the difficulty of capturing pronunciation changes in low-resource ASR. An improved GRU network was added to the back end of the model to enhance the alignment of phone frame states, and a multi-head attention was introduced, combining rough and fine-grained features of audio and spectrum to highlight differences in resonance peaks. The experimental results show that adding an improved GRU can reduce WER by 1.92%, 0.38%, and SER by 5.6% and 4.4%, respectively, by adding a PDP module. Meanwhile, after introducing multi-head attention, the results showed that adding a PDP module reduced WER by 2.33%, 0.45%, and SER by 6.0% and 4.8%, respectively [11]. Wang et al. proposed a two-channel SS fusion capsule generation adversarial net for HSI classification. To further improve the classification performance, an SS channel fusion model was constructed to synthesize and switch the feature information of different channels, thus improving the accuracy and robustness of the overall classification performance. The experimental performance shows that the proposed model can effectively improve the classification accuracy and performance [12]. Rezgui and Marks explored the factors that influence the information security awareness of higher education staff, including information system decision-makers, in developing countries. The study showed that factors such as responsibility, cultural assumptions and beliefs, and social conditions usually influence the behavior and work attitudes of university staff, especially information security awareness [13].

To sum up, the intelligence and informatization of life make people communicate with each other more and more frequently, and the voice data generated is immeasurable. In the field of information security-oriented and intelligent speech, rapidly locating the identity of the speaker and speaker identification of the channel is a critical issue. This research proposes a vocal recognition technique based on channel adversarial training for such a problem, which has practical implications for intelligent applications in the field of information security.

3. Recognition Method of Speech Recognition Technology based on Channel Adversarial Training

Since individual voice is always stored and transmitted on different hardware and applications, it is important to identify speakers across channels. However, in the deep learning processing method of intelligent speech, the end-to-end speaker representation has a good prospect. Therefore, this chapter proposes the identification method of voice print recognition technology based on the adaptation of the channel domain.

3.1. A voice information database of the speaker recognition system

In real life, the scale of audio data is very large, and the centralized storage method of the server is unable, to meet the demand of massive data storage and management [14]. Therefore, a distributed storage scheme is used in the system, which mainly concentrates the disk space on the machines within the network into a virtual storage space, and even though the physical storage is distributed everywhere, it can still be used as a virtual overall space to provide storage services to the outside world [15]. The storage and management of massive audio data are shown in Figure 1.

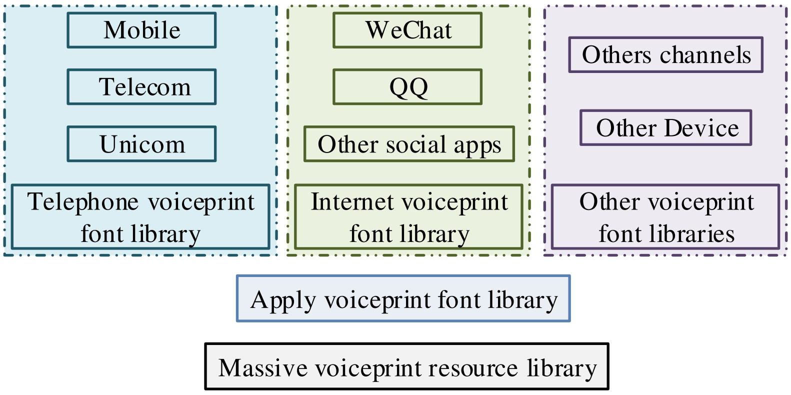

Mobile
Telecom
Unicom
Telephone voiceprint font library
WeChat
QQ
Other social apps
Internet voiceprint font library
Others channels
Other Device
Other voiceprint font libraries
Apply voiceprint font library
Massive voiceprint resource library

Figure 1. Schematic diagram of voice storage database.

Figure 1 is a massive voice data resource database, which is mainly composed of telephone voice, Internet voice, and other voice, the three sub-speaker voice databases. In the face of the storage of massive speaker representations, distributed storage is the storage of externally provided interfaces for file upload, file download, file modification and file deletion [16]. Distributed storage is to create the voice print sublibrary based on the gender, age, and other information of the speaker, which is the most concise and effective method to build the voice print word database, as shown in Figure 2.

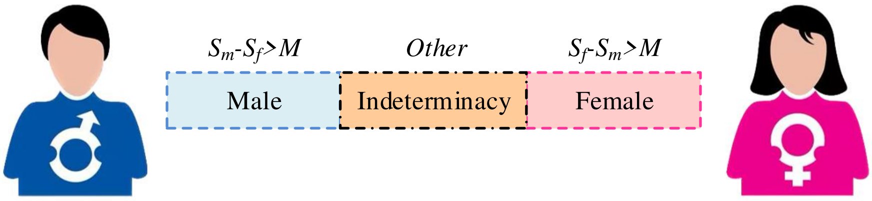


Figure 2. Schematic diagram of voiceprint sub library division.

In Figure 2, an error tolerance mechanism is added to the word bank-building process. S_m represents the male confidence score predicted by the gender recognition model, S_f represents the female confidence score. When S_f - S_m is used, the speaker is labeled as female, when S_m - S_f is used, the speaker is labeled as male, and when both conditions are not met, it is labeled as uncertain. The machine automatically gives a gender label to the speaker in the library, and when retrieving the test speech, it can go directly to the corresponding word bank after obtaining the information [17]. However, gender-based identification usually requires the speaker-related and irrelevant judgment of the category information after clustering. Among them, the most common application mode is the text-irrelevant speaker identification and text-related speaker identification system. Its structural process is shown in Figure 3.

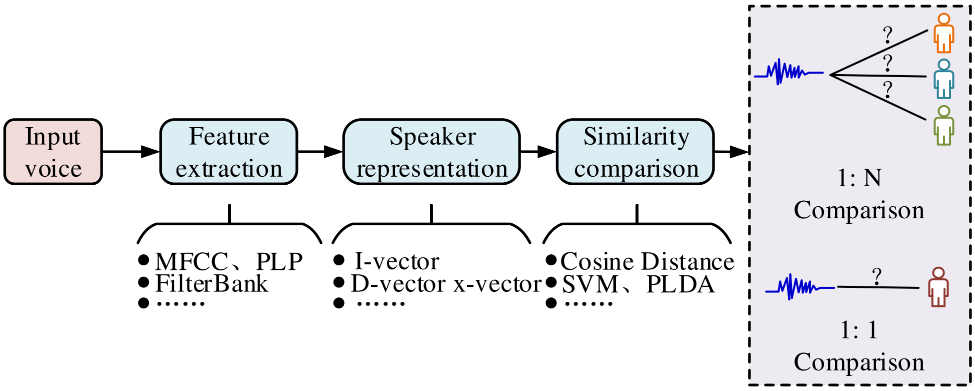


Figure 3. Voiceprint recognition flowchart.

A flow chart of vocal recognition based on multiple methods under text-independent speaker recognition and text-dependent speaker confirmation systems is shown in Figure 3. In practice, there is also often a scene and channel mismatch between the registration and test datasets, which is less effective for speaker recognition [18]. Therefore, for a speech recognition system, a detection cost function (DCF) is given as an evaluation metric using the expression shown in Equation (1).

$$DCF = C_{fa} \cdot FA \cdot P_{\mathrm{Im}\,p} + C_{fr} \cdot FR \cdot P_{Tar} \qquad (1)$$

In Equation (1), FA is the false acceptance rate, FR is the false rejection rate, C_{fa} denotes the false rejection cost, C_{fr} denotes the false reception cost, $P_{\mathrm{Im}\,p}$ denotes the non-target speaker prior probability, and P_{Tar} denotes the target speaker prior probability. The values of these four variables can be adjusted according to different application scenarios. And in order to solve the scenario channel mismatch, the $i-vector$ back-end system is used to compensate the channel. PLDA is a

doi:10.20944/preprints202307.1591.v1

channel compensation algorithm that can decouple the guaranteed information more thoroughly, and its calculation formula is shown in Equation (2).

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \varepsilon_{ij} \qquad (2)$$

In Equation (2), x_{ij} denotes the $i-vector$ extracted from the j speech of the i speaker, μ denotes the overall mean of the training data $i-vector$, F denotes the speaker information space, G denotes the channel space, ε_{ij} denotes the residual term, h_i denotes the speaker space coordinates, and w_{ij} denotes the channel space coordinates. PLDA is tested either by extracting h_i to calculate the cosine distance score or by directly calculating the likelihood generated from h_i, as defined in Equation (3).

$$score = \log \frac{p(x_1, x_2 | H_S)}{p(x_1, | H_d) p(x_2, | H_d)} \qquad (3)$$

In Equation (3), x_1 and x_2 denote the $i-vector$ of two voices, H_s and H_d denote the same speaker and different speakers, $p(x_1, x_2 | H_S)$ denotes the x_1 and x_2 likelihood functions of the same speaker, and $score$ denotes the final score. Deep learning approaches have become the main technique for speaker recognition, whereas the end-to-end approach has become one of the mainstream approaches in the field of pattern recognition. The advantage of the end-to-end framework is that it can directly have the optimization of the target, and in the speech separation task, Scale-Invariant Source-to-Noise Ratio (SI-SNR) is used as the metric, which is calculated as shown in Equation (4).

$$SI - SNR = 10 \log_{10} \frac{\| St\arg et \|^2}{\left\| \hat{S} - St\arg et \right\|^2} \qquad (4)$$

In Equation (4), $St\arg et = \frac{< \hat{S}, S > S}{\|S\|^2}$, $\hat{S}$ and S are the estimated and target values, respectively. For multiple outputs, the overall loss function is usually based on the displacement-invariant training learning target, which is calculated as shown in Equation (5).

$$L = \frac{1}{N} \sum_{i=1}^{N} l(\hat{S}_i - S_\phi) \qquad (5)$$

In Equation (5), N denotes the number of speakers, l denotes the error between the network output and the target, $\hat{S}_i$ denotes the first i predicted voice, and S_ϕ denotes the reference voice with the alignment order ϕ that minimizes the training target L . The end-to-end speaker characterization using deep neural networks to compensate for channel mismatch is even more widely used.

3.2. Voice recognition based on channel adversarial training

As deep learning becomes a major technique in the field of intelligent speech and is popularly used in speaker recognition, end-to-end speaker-based characterization has a good prospect [19]. However, for the training data collected from different channels, the convolutional neural network cannot directly model the speaker information between different channels. In this study, based on an unsupervised domain adaptive approach, we propose a model by channel adversarial training (CAT), which relies only on the speaker's speech data under each channel and does not require the same speaker's speech data under different channels. The model structure is shown in Figure 4.

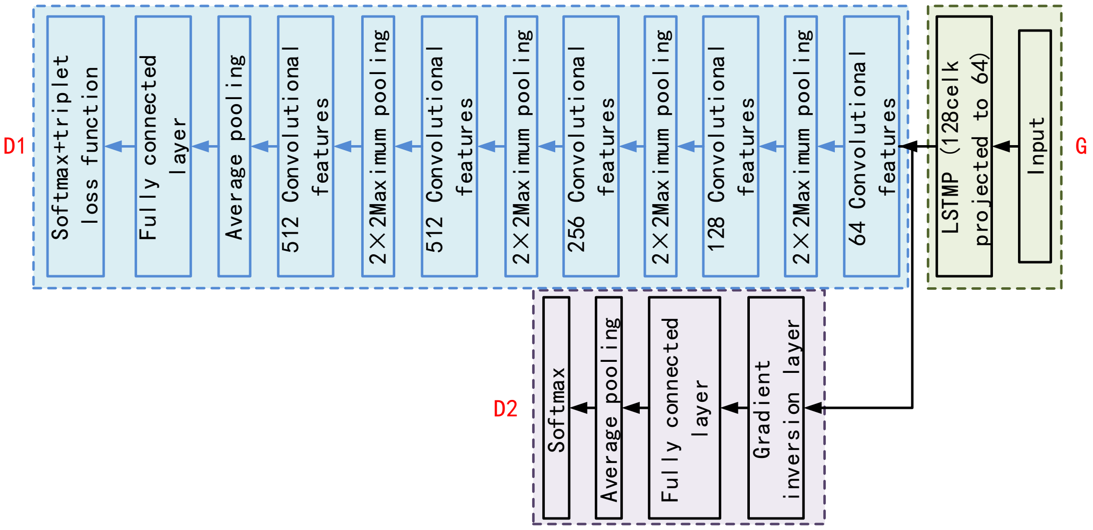

D1
Softmax+triplet loss function
Fully connected layer
Average pooling
512 Convolutional features
2×2Maximum pooling
512 Convolutional features
2×2Maximum pooling
256 Convolutional features
2×2Maximum pooling
128 Convolutional features
2×2Maximum pooling
64 Convolutional features
LSTMP (128celk projected to 64)
Input
G
D2
Softmax
Average pooling
Fully connected layer
Gradient inversion layer

Figure 4. Schematic diagram of text independent speaker recognition network structure based on channel confrontation.

Figure 4 shows the structure of the speaker recognition network based on channel adversarial training with a baseline CNN model. Five convolutional layers are included, and the input layers piece together the features of the same person to form a feature map. The overall loss function of the model consists of Softmax and Triplet loss functions together. Its calculation formula is shown in Equation (6).

$$L_s = -\sum_{i=1}^{M} \log \frac{e^{W_{yi}^T x^i + b_{yi}}}{\sum_{j=1}^{N} e^{W_j^T x^i + b_j}} \qquad (6)$$

In Equation (6), x^i is represented as the representation of the i speaker, which belongs to the speaker y^i. w^j denotes the last fully connected column j and b is the bias term. the size of the minibatch is M and the number of speakers is N. The Triplet loss function is defined as shown in Equation (7).

$$L_T = \sum_{i=1}^{M} \max(0, D(x^i, x^n) + \delta - D(x^i, x^p)) \qquad (7)$$

In Equation (7), the two samples in (x^i, x^p) are from the same speaker, while (x^i, x^n) is from a different speaker and x^i is the anchor sample of the triad. $D(x^i, x^p)$ denotes the cosine distance between the two input vectors, and the two loss functions are superimposed and adjusted by the weights α, which are calculated as shown in Equation (8).

$$L = L_s + \alpha L_T \qquad (8)$$

In the face of excessively long sentences, the long sentences are segmented into multiple short segments by using sliding windows that do not overlap each other. And then the sentence-level speaker representation vector is obtained by averaging pooling. The model is decomposed into three different neural networks, including a feature extractor G, a speaker label classifier $D1$ and a channel label classifier $D2$, and the expressions are shown in Equation (9).

$$\begin{cases} G = f_G(x, \theta_G) \\ D1 = f_{D1}(g, \theta_{D1}) \\ D2 = f_{D2}(g, \theta_{D2}) \end{cases} \qquad (9)$$

In Equation (9), θ_G, θ_{D1}, and θ_{D2} represent the parameters of each network, respectively, by using a gradient inversion layer and optimizing θ_G to minimize the speaker prediction loss and maximize the channel classification loss. The hyperparameter β is used to balance the $D1$ and $D2$ losses in the backpropagation process to obtain a representation of channel invariance and speaker differentiation. The overall optimized loss function is a combination of the $D1$ and $D2$ loss functions. $D1$ The loss function is defined as shown in Equation (10).

$$L_{D1} = L_s + \alpha L_T \tag{10}$$

In Equation (10), the overall loss function consists of Softmax and Triplet loss functions together and is adjusted by the weights α, $D2$ The loss function is defined as shown in Equation (11).

$$L_{D2} = -\sum_{i=1}^{M} \log \frac{e^{W_{di}^T x^i + b_{di}}}{\sum_{j=1}^{K} e^{W_j^T x^i + b_j}} \tag{11}$$

In Equation (11), denotes the representation of the x^i i th speaker, which belongs to the speaker d^i. w^j denotes the last fully connected column of j and b is the bias term. the size of the minibatch is M and the number of speakers is N. The whole CAT framework is optimized using the stochastic gradient descent method. The optimal parameters are obtained from the optimization process of the following equation as shown in Equation (12).

$$\begin{cases} \theta_G = \theta_G - l * (\dfrac{\partial L_{D1}}{\partial \theta_G} - \beta * \dfrac{\partial L_{D2}}{\partial \theta_G}) \\ \theta_{D1} = \theta_{D1} - l * (\dfrac{\partial L_{D1}}{\partial \theta_{D1}}) \\ \theta_{D2} = \theta_{D2} - l * (\dfrac{\partial L_{D2}}{\partial \theta_{D2}}) \end{cases} \tag{12}$$

In Equation (12), speaker representations with channel invariance, as well as speaker differentiation, are extracted directly from the network after training with the above parameters. This method can map two different channels within a common subspace, train processes with channel invariance, as well as obtaining speaker representations of speaker discriminability. Thus eliminating channel differences and improving the performance of speaker recognition. This framework alleviates channel mismatch, through channel antagonism training, overcomes channel divergence, and improves applicability in real scenarios. Since most of the massive voice data comes from telephone, Internet audio and video, and other APP applications, the schematic diagram of the speaker recognition system in the field of information security is shown in Figure 5.

doi:10.20944/preprints202307.1591.v1

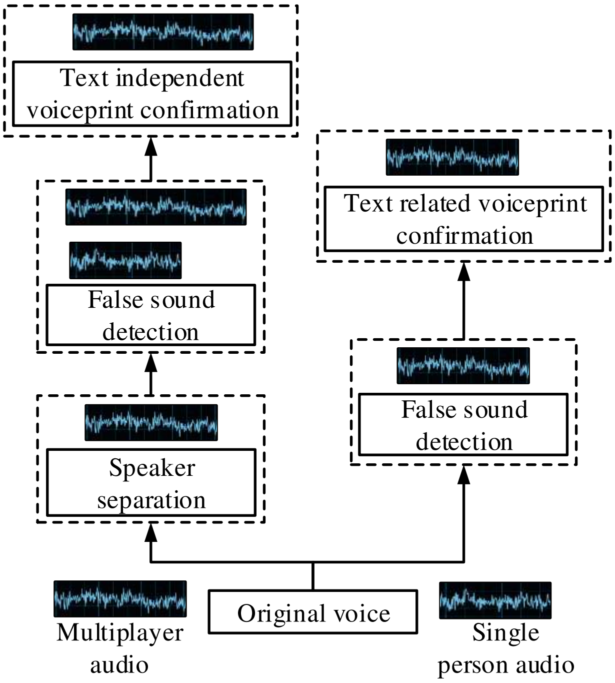


Figure 5. Schematic diagram of a speaker recognition system for the field of information security.

Figure 5 shows the speaker identification system in the field of information security. The system needs to have the data storage and fast retrieval ability of the speaker. It mainly aims at the separation of the speaker, text-related speaker confirmation, and false detection. Where the resulting separation error rate refers to the percentage of the length of the entire effective speech length and is defined as shown in Equation (13).

$$DER = FALSE + MISS + SER \qquad (13)$$

In Equation (13), $FALSE$ denotes the false alarm rate of valid speech detection, $MISS$ denotes the miss detection rate of valid speech detection, and SER denotes the speaker-to-speaker classification error rate. In the false voice detection task, an acoustic feature CQCC based on Constant Q Transform (CQT) combined with cepstral generation is commonly used. it first performs the CQT transform on the speech signal as shown in Equation (14).

$$X^{CQ}(k,n) = \sum_{j=n-|N_k/2|}^{j=n+|N_k/2|} x(j)a_k^*(j-n+N_k/2) \qquad (14)$$

The system extracts the speaker representation for each speech data, extract the speaker information, and store and retrieve it. The control layer includes the functions of task scheduling, data management, and plug-in management, and the plug-in is the process as shown in Figure 6.

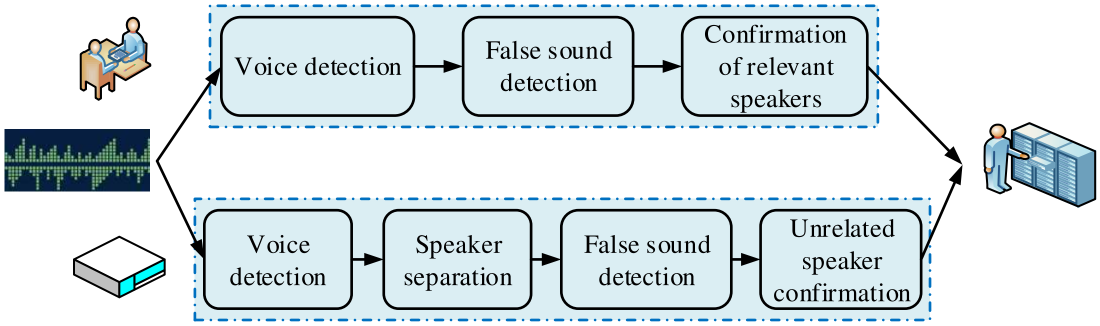
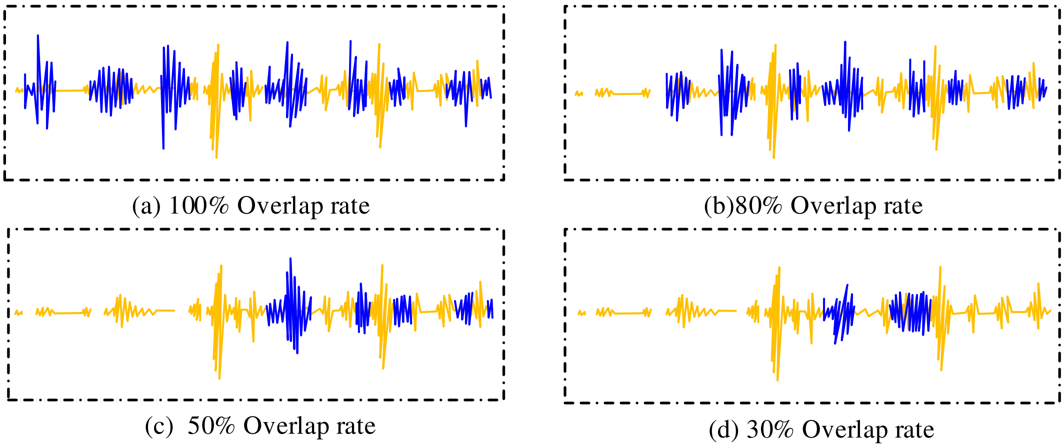


Figure 6. Plug-in process diagram.

Plug-in management is the voice detection, false sound detection, speaker separation, and text-related speaker recognition as a separate plug-in to quickly synthesize the application market demand module [20]. In addition, fast switching to application scenarios is achieved through the dynamic combinability of functional layers to enable plug-in management of the application market. Plug-in management enables not only fast configuration of application scenarios after selection but also fast switching of multiple application scenarios in a system.

4. Research on the Application of Speech Recognition Technology based on Channel Adversarial Training in the Field of Information Security

This chapter collects audio from several real-life scenarios, including conversations of phone calls, meetings, multi-person chats, and other types. While current speech separation techniques perform better in the case of two people speaking, they are less effective in the case of multiple people speaking. This research addresses the enhancement of the speech separation technique in the case of multiple speakers. By simulating the richness of speech in real conversation scenarios, data with different speech overlap rates are added, and the data are shown in Figure 7.


(a) 100% Overlap rate

(b)80% Overlap rate

(c) 50% Overlap rate

(d) 30% Overlap rate

Figure 7. Schematic diagram of overlapping data with different speech overlap rates.

As can be seen from Figure 7, the four different groups of speech overlap rates present different speech styles, and speech separation is a good separation scheme for the task of separating speech overlap segments. Additive sparse mixing of speech with different overlap rates is performed, which is useful for the study to improve the generalization to different speakers. It is obvious from the different speech overlap rates that increasing the size of the training data gradually improves the overall performance. Adding them to the model training further improves the DER from 12.91% to 12.78%. However, if the channel adversarial training strategy is not used, its experimental results are changed in this study. The comparison of the results by removing the D2 module from the CAT framework in this study is shown in Table 1.

Table 1. Comparison of EER (%) results between CAT method and other methods on the development set.

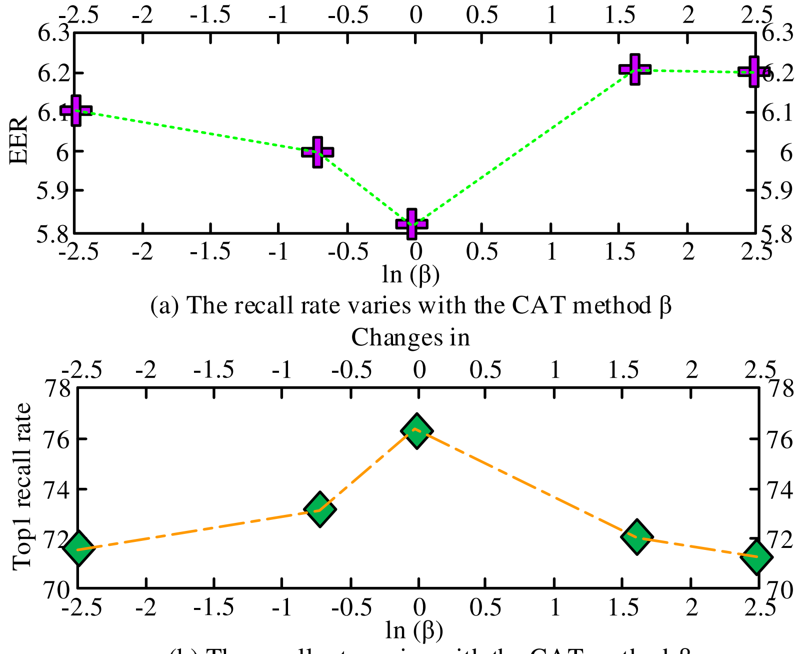
System	EER			
	First training	Second training	Third training	Fourth training
I-vector	8.71%	8.82%	8.62%	8.83%
CNN	6.23%	6.13%	6.24%	6.34%
CAT without D2	6.42%	6.41%	6.57%	6.42%
CAT	5.81%	5.91%	5.70%	5.83%

As can be seen from Table 1, the removal of the D2 module in the CAT framework, which is the channel classifier, enables the model complexity to be compared with no channel adversarial training strategy under this approach. The EER results of the CAT method are compared with those of the I-vector, CNN, and CAT with the D2 module removed, and the EER results of the I-vector method are around 8.5%, the EER results of the CNN and CAT without D2 methods are around 6.2%, and the EER results of the CAT method are around 5.8%. After conducting a comparison of the four sets of results, it is clear that simply increasing the feature extractor G does not improve the performance. This shows the effectiveness of the adversarial strategy proposed in this study. To further validate the performance of the speaker recognition task on large-scale data, the performance metric of the test set is TopN recall, and the CAT method is compared with I-vector, CNN, and CAT method with the D2 module removed for TopN recall results, which are shown in Table 2.

Table 2. Comparison of TopN (%) results between CAT method and other methods on the test set.

System	Recall		
	Top1	Top5	Top10
I-vector	57.11%	66.22%	70.13%
CNN	69.21%	77.23%	79.91%
CAT without D2	68.92%	77.81%	79.84%
CAT	76.21%	83.15%	84.92%

As can be seen from Table 2, the speaker recognition performance achieved a relative 22.3% (6.5% in absolute value) improvement in Top1 recall compared to the baseline after the CAT method. In particular, the TopN recall on the test set for the CAT method compared with the I-vector, CNN, and CAT with the D2 module removed, the recall under the I-vector method improved by 27%. The recall under both the CNN and CAT without D2 methods improved by about 10%, and the recall under the CAT method improved over the other methods by 6.5%. This shows that the test set and the development set are consistent in terms of data performance, further demonstrating the effectiveness of the CAT method. In order to balance the two loss functions under the CAT framework, the variation of EER and Top1 metrics with the parameter β was investigated using the hyperparameter β, and the results are shown in Figure 8.

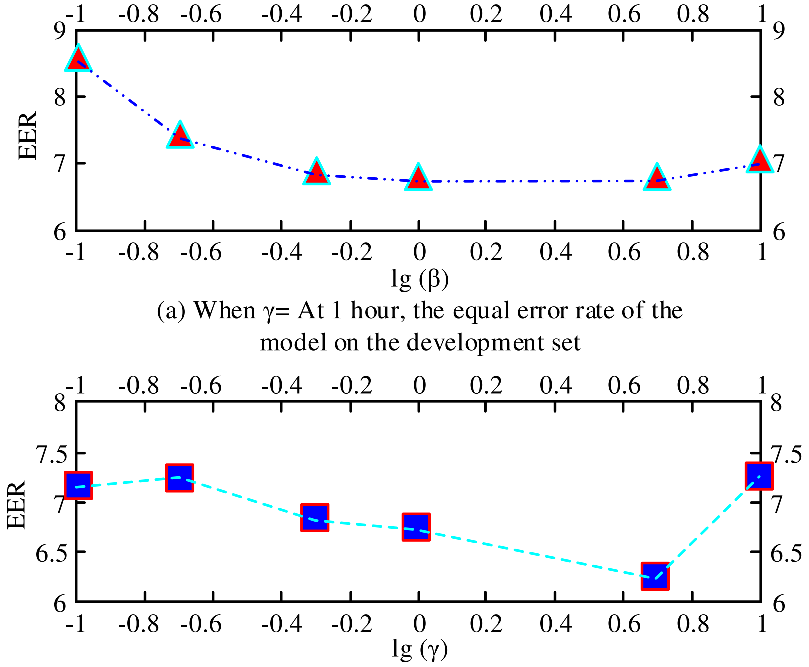


(a) The recall rate varies with the CAT method β Changes in

(b) The recall rate varies with the CAT method β Changes in

Figure 8. The recall rate of EER and Top1 under CAT method varies with β Schematic diagram of changes in.

As can be seen from Figure 8, when the hyperparameter β is set to 0, the EER indicator is 5.8, which presents the lowest value, and when the hyperparameter β is 1.68, the EER indicator is 6.23, which is the highest value. However, when the hyperparameter β is set to 0, the Top1 indicator is 76.69, which presents the highest value, and when the hyperparameter is 2.48, the Top1 indicator is 71.73, which is the lowest value. It can be seen that the lowest EER indicator and the highest Top1 indicator both appear when the hyperparameter β is set to 0. In addition, two different loss function weights of the hyperparameters β and γ were used to test the performance of the model, and the effects of the two hyperparameters on the EER are shown in Figure 9.

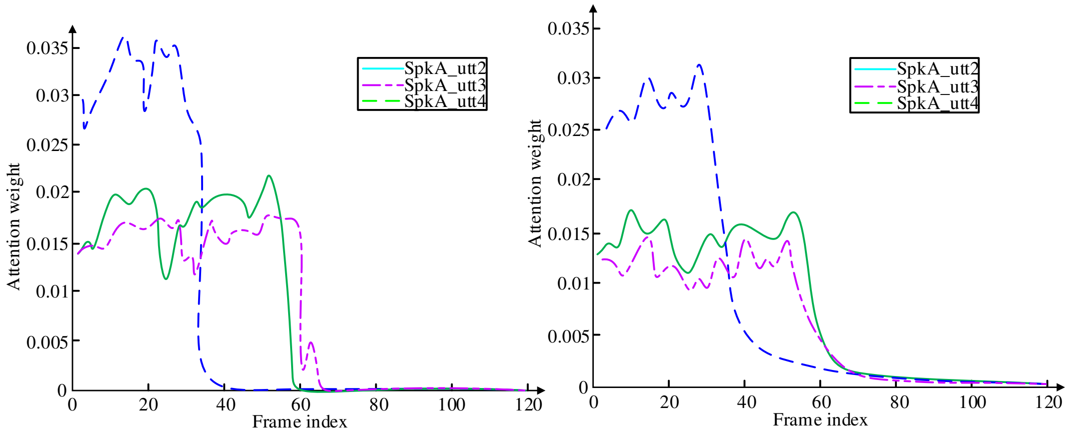


(a) When γ= At 1 hour, the equal error rate of the model on the development set

(b) When β= At 1 hour, the equal error rate of the model on the development set

Figure 9. Schematic diagram of EER corresponding to different weight losses.

As can be seen from Figure 9, when γ is initialized to 1, the effect of β on the EER of the development set shows a decreasing trend, and when β is set to -1, the EER indicator has the highest value of 8.76, and when β is set to 0, the EER indicator has the lowest value of 6.71. When β is initialized to 1, the effect of γ on the EER of the development set shows a decreasing trend, and when γ is set to 1, the EER indicator has the highest value of 7.48, and when γ is set to 0.7, the EER indicator has the lowest value of 6.31. When γ is set to 0.7, the EER indicator is the lowest value of 6.31. Thus, it can be seen that when β is set to 0 and is set to 5, the EER indicator reaches the lowest EER indicator of 6.12 in the channel adversarial training method. In addition, the bidirectional attention mechanism designed for end-to-end text-irrelevant speaker recognition considering the text-irrelevant speaker recognition task and thinking about the selective auditory attention of the human brain is used. The data of attention weights corresponding to registered speech and test speech are analyzed under the CNN model and BaCNN model. The results of their analysis are shown in Figure 10.



(a) Attention Weight Curve in CNN Model (b) Attention Weight Curve in BaCNN Model

Figure 10. Illustrative diagram of attention weight.

From Figure 10, where SpkA_utt2 *and SpkA_utt3 are the speech* 2 and speech 3 of the exploring speaker and *SpkA_utt4* is the speech 4 of the speaker for the distribution of attention weights when testing the speech. The horizontal coordinate is the frame index and the vertical coordinate is the attention weight coefficient. the attention weight of *SpkA_utt2 in the* CNN model is highly similar to the corresponding *SpkA_utt3, which can be* large when the test speech is from a different speaker. the attention weight of *SpkA_utt2* in the BaCNN model is 35 frames ahead of *SpkA_utt4.* The attention weights of the registered speech differ depending on the test speaker, which indicates that the representation of interactive speech from different speakers is indeed learned in the BaCNN model.

5. Conclusion

About speaker voice recognition technology has a very important application value in the field of information security. With the recognition needs of telephone, Internet, and various APP applications, for the speaker voice recognition technology for the information security field. This research proposes a vocal recognition technology based on channel confrontation training for the task of text-independent speaker recognition and text-related speaker confirmation for massive audio and video in the field of information security. Additive sparse mixing of speech with different overlap rates is performed. It is evident from the different speech overlap rates that increasing the size of the training data gradually improves the overall performance. Adding them to the model training further increases the DER from 12.91% to 12.78%. The lowest EER metric and the highest Top1 metric both occur when the hyperparameter is set to 0. The recall rates under both CNN and CAT without D2 methods improved by about 10%, and the recall rates under the CAT method both improved by 6.5% over the other methods. Setting to 0 and to 5, the EER metric reached the lowest EER metric of 6.12 in the channel adversarial training method. In addition, the bidirectional attention mechanism was designed by considering the text-independent speaker recognition task and thinking about the selective auditory attention of the human brain. the attention weights of SpkA_utt2 in the CNN model are highly similar to the corresponding SpkA_utt3. SpkA_utt2 in the BaCNN model is 35 frames ahead of SpkA_utt4 in the BaCNN model, which does learn the representations of interactive speech of different speakers. In conclusion, the vocal recognition technique based on channel adversarial training obtains the speaker's representations, improves the channel mismatch problem, and enhances the vocal recognition effect.

Funding: This work was supported by The National Key Research and Development Program of China (NO. 2020YFC0833400)

References

1. Feng Y. Make the rocket intelligent at IOT edge: Stepwise GAN for anomaly detection of LRE with multi-source Fusion. IEEE Internet of Things Journal, 2021, 9(4): 35-49.
2. Li Q, Qu H, Liu Z, Zhou N, Sun W, Sigg S, Li J. AF-DCGAN: Amplitude feature deep convolutional GAN for fingerprint construction in indoor localization systems.Networking and Internet Architecture, 2021, 5(3): 468-480.
3. Shen X, Jiang H, Liu D, Yang K, Deng F, Lui J C S, Luo J. PupilRec: Leveraging pupil morphology for recommending on smartphones. IEEE Internet of Things Journal, 2022, 9(17), 15538-15553. doi: 10.1109/JIOT.2022.3181607
4. Yan L, Shi Y, Wei M, Wu Y. Multi-feature fusing local directional ternary pattern for facial expressions signal recognition based on video communication system. Alexandria Engineering Journal, 2023, 63, 307-320. doi: https://doi.org/10.1016/j.aej.2022.08.003
5. Khdier H Y, Jasim W M, Aliesawi S A. Deep learning algorithms based voiceprint recognition system in noisy environment. Journal of Physics Conference Series, 2021, 1804(1): 12-42.
6. Sun W Z, Wang J S, Zheng B W, Zhong-Feng L. A novel convolutional neural network voiceprint recognition method based on improved pooling method and dropout idea. IAENG Internaitonal Journal of Computer Science, 2021, 48(2): 202-212.
7. Hong Z, Yue L, Wang W, Zeng X. Research on end-to-end voiceprint recognition model based on convolutional neural network. Journal of Web Engineering, 2021, 20(5): 1573-1586.
8. Zhang T, Waqas M, Shen H, Liu Z, Zhang X, Li Y, Halim Z, Chen S. A neural network architecture optimizer based on DARTS and generative adversarial learning. Information Sciences: An International Journal, 2021, 581(20): 448-468.
9. Yu H K, Nam S H, Hong S B, Park K R. GRA-GAN: Generative adversarial network for image style transfer of gender, race, and age. Expert Systems with Applications, 2022, 198(6): 2-20.
10. Wang C, Luo D, Liu Y, Xu B, Zhou Y. Near-surface pedestrian detection method based on deep learning for UAVs in low illumination environments. Optical Engineering, 2022, 61(2): 2-19.
11. Zhu W, Jin H, Chen J, Luo L, Lu Q, Li A.A hybrid acoustic model based on PDP coding for resolving articulation differences in low-resource speech recognition. Applied Acoustics, 2022, 192(4): 2-11.
12. Wang J, Guo S, Huang R, Li L, Zhang X, Jiao L. Dual-channel capsule generation adversarial network for hyperspectral image classification. Transactions on Geoscience and Remote Sensing, 2021, 60(99): 2-16.
13. Shan S, Liu J, Dun Y. Prospect of voiceprint recognition based on deep learning. Journal of Physics: Conference Series, 2021, 18(1): 12-46.
14. Ji H, Lei X, Xu Q, Huang C, Ye T, Yuan S. Research on characteristics of acoustic signal of typical partial discharge models. Global Energy Interconnection, 2022, 5(1): 118-130.
15. Cai R, Wang Q, Hou Y, Liu H. Event monitoring of transformer discharge sounds based on voiceprint. Journal of Physics: Conference Series, 2021, 2066(1): 66-67.
16. Qian W, Xu Y, Zuo W, Li H. Self-sparse generative adversarial networks. CAAI Artificial Intelligence Research, 2022, 1(1): 68-78.
17. Kim J I, Gang H S, Pyun J Y, Goo-Rak K. Implementation of QR code recognition technology using smartphone camera for indoor positioning. Energies, 2021, 14(2): 59-63.
18. Zhu K, Ma H, Wang J, Yu C. Optimization research on abnormal diagnosis of transformer voiceprint recognition based on improved wasserstein GAN. Journal of Physics Conference Series, 2021, 17(4): 12-67.
19. Yang Y, Song X. Research on face intelligent perception technology integrating deep learning under different illumination intensities. Journal of Computational and Cognitive Engineering, 2022, 1(1): 32-36.
20. Amin S N, Shivakumara P, Jun T X. An augmented reality-based approach for designing interactive food menu of restaurant using android. Artificial Intelligence and Applications. 2023, 1(1): 26-34.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.