International Journal of Modern Achievement in Science, Engineering and Technology (IJSET)
ISSN 3023-459X
DOI: 10.63053/ijset.63

A review of speaker identification methods with emphasis on new approaches

Ali Azam[1],**Masoumeh Shafieian**[2]
1.Master student of Islamic Republic of Iran Broadcasting University,Tehran,Iran
2.Member of the Faculty of Broadcasting University of the Islamic Republic of Iran

ARTICLE INFO

Keywords:
Speaker Identification, Deep Learning, Speech Processing

ABSTRACT

Speaker identification is one of the important and practical challenges in the field of speech processing, which plays a significant role in security, voice authentication, and intelligent systems. This article examines the new methods of speaker identification and analyzes the recent developments in this field. The main focus of the paper is on the introduction and analysis of modern deep learning techniques, including convolutional neural networks (CNN) and hybrid models capable of extracting and analyzing more complex features of the speech signal. These methods have many applications in identifying the speaker independent of the text and in noisy or non-ideal conditions. Finally, remaining challenges, existing limitations, and future research directions for the development of more accurate and stable systems are reviewed. This study can help researchers and developers in a better direction in this field.

Introduction

Speaker identification is one of the critical and complex issues in speech processing, which has attracted more attention in recent decades, especially with significant advances in machine learning and signal processing. It plays a prominent role in security applications, voice authentication, access control systems, customer service and voice assistants. Speaker identification is related to the process of identifying a person's identity through the analysis of the unique acoustic characteristics of the speech signal, which is especially useful in identifying people in different conditions such as noisy environments or with limited audio data.

The speaker recognition process usually involves several steps, which generally include four main steps: speech data preprocessing, feature extraction, speaker modeling, and classification. In the pre-processing stage, the speech signal is entered and various steps such as framing, pre-emphasis and windowing are performed on it. Then, in the feature extraction stage, certain features are extracted from the audio signal that are able to effectively simulate the identity of the speaker. Common features are Mel-Frequency Capstral Coefficients (MFCC) and Linear Prediction Coefficients. In the speaker modeling stage, these features are given to different models such as Gaussian mixture models, hidden Markov models or neural networks to simulate the speaker. Finally, using the constructed models, the speakers are separated and identified.

Speaker recognition systems are usually divided into two main categories: text-dependent and text-independent systems. In text-dependent systems, the speaker must utter a specific phrase, which limits the variety of audio inputs. While in text-independent systems, the speaker can say anything and the system must be able to recognize his identity in any situation. These types of systems are especially important in applications that require high accuracy and flexibility, such as authentication systems, access control to sensitive resources, and information security.

In recent years, due to the significant advances in the fields of deep learning, new approaches for speaker identification have been proposed, which are able to more accurately and quickly identify the unique features of speakers from audio signals. These methods have not only improved the recognition accuracy, but also are able to work with audio data in different conditions, including noisy audio data and low-quality signals.

The purpose of this article is to collect and review different methods based on artificial intelligence to identify the speaker. In this article, the comparison of traditional and new techniques in this field is discussed, as well as challenges and future research directions are examined. In particular, methods used in the fields of machine learning and signal processing to improve speaker recognition systems, such as techniques based on deep neural networks, will be reviewed.

Classification and improvements of speaker identification methods

Speaker identification is one of the most important areas of speech processing, which involves extracting unique acoustic features from speech signals to identify the speaker. According to the developments in this field, speaker recognition systems are divided into two main categories: text-dependent systems and text-independent systems.

Text-based systems

In these types of systems, the speaker must express a specific phrase. These systems are mostly used in applications such as password recognition or identity verification based on a predetermined phrase. Models used for this type of speaker identification mainly focus on features such as Mel-Frequency Capstral Coefficients (MFCC) and linear prediction coefficients. These systems work well in certain situations and with predictable sentences, but face limitations when dealing with sentence variety or unexpected speech situations.

Text independent systems

These systems are particularly useful for applications such as voice authentication and access control, since the speaker can say anything and the system must identify him or her. In these systems, acoustic features are continuously extracted from the speech signal and the system must find certain patterns of the speaker's speech. Text-independent systems require more sophisticated algorithms that can take advantage of general speech features and unique speaker characteristics in different situations (such as noise or changes in speech).

Traditional and parametric methods

In the past, traditional methods such as hidden Markov models (HMM) and Gaussian mixture models (GMM) were used for speaker recognition. These methods mostly relied on features such as Capstral coefficients or linear prediction coefficients. These methods usually work well in controlled conditions, but their performance degrades in more complex conditions such as noise or changes in the speaker's speaking conditions.

Methods based on deep learning

With recent advances in deep learning, new speaker recognition methods have been introduced that are able to detect more complex features and hidden patterns in speech data. Among these methods are convolutional neural networks (CNN), recurrent neural networks (RNN) and deep neural networks (DNN). Using features such as spectrograms and spectrograms with Mel scale, these models are able to extract more complex features from speech and improve speaker identification in more complex conditions such as noise or speech.

Using models with self-attention mechanism

One of the new developments in this field is the use of models based on the attention mechanism (Self-Attention) such as transformers, which allows the model to focus on different parts of the input and simulate the relationships between different parts of the speech signal. These models are very effective in speaker recognition due to their high capabilities in modeling sequences and processing complex temporal data.

Challenges and future methods

Despite significant advances in speaker recognition techniques, there are still challenges such as background noise, changes in speaker speaking conditions, and similar speakers that need to be addressed in developing more accurate and scalable systems. The future of speaker recognition depends on the use of a combination of deep learning and new machine learning techniques that can improve the performance of systems in more complex situations.

Evaluation criteria:

Accuracy evaluation criteria:

One of the key evaluation criteria in speaker identification is accuracy. Accuracy represents the accuracy of the model's predictions and shows how many of the predicted samples are correctly attributed to the original speaker. This criterion is especially important in speaker recognition systems, as inaccuracies in recognition may lead to poor performance in real-world applications, including in authentication systems and voice assistants.

The accuracy criterion is defined according to equation 1:

Accuracy = (1)

Classification Error Rate (CER) evaluation criterion:

Classification Error Rate (CER) is a measure used to measure the amount of model errors in speaker identification. CER specifically represents the total percentage of samples that are incorrectly classified. This measure can be particularly useful in situations where error reduction is important.

CER criterion is defined according to equation 2:

CER = 1 - Accuracy (2)

In this regard:

Accuracy is a measure that shows what percentage of all samples are correctly identified.

CER is calculated inversely from Accuracy and represents the percentage of samples that the model predicted incorrectly.

CER can help analyze the performance of the model and evaluate its weak points in different conditions. Using CER along with other metrics such as accuracy, sensitivity, and specificity can provide a more comprehensive picture of speaker recognition model performance.

Top-1 Accuracy evaluation criteria:

Top-1 Accuracy is a measure used to evaluate the performance of the model in correctly identifying the speaker. This measure indicates the percentage of predictions that the model has made correctly, so that the correct answer is at the top of the list of model predictions.

The Top-1 Accuracy criterion is defined according to equation 3:

$$\text{Top-1 Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (3)$$

In this regard:
Number of correct predictions: The number of times the prediction with the highest probability of the model matched the correct answer.
Total Number of Predictions: The total number of predictions made by the model.

Top-5 Accuracy Evaluation Criteria:
Top-5 Accuracy is a measure used to evaluate the performance of the model in correctly identifying the speaker and indicates the percentage of predictions that include the correct answer among the five predictions with the highest probability of the model. This criterion is useful, especially when the correct speaker choice may be among several possible choices.
The Top-5 Accuracy criterion is defined according to equation 4:

$$\text{Top-1 Accuracy} = \frac{\text{The number of predictions where the correct answer is among the five predictions with the highest probability.}}{\text{The total number of predictions}} \quad (4)$$

In this regard:

Number of predictions where the correct answer is among the 5 predictions with the highest probability: The number of times the correct answer is among the five predictions with the highest probability of the model.

Total Number of Predictions: The total number of predictions made by the model.

A review of previous methods

In 2015, Al-Medid et al. (Almaadeed, Aggoun, and Amira 2015) designed and implemented a text-independent speaker recognition system using wavelet transform analysis and neural networks to increase the speed and accuracy of classification. In this system, first, in order to decompose the given speech signal into a group of smaller signals at different levels and analyze each component of the speech signal at different frequencies with different resolutions, wavelet transformation was applied, and in the second step, the distinctive features of the whole signal The speech was extracted .

In 2015, Daqrouq and Tutunji (2015) proposed a new feature extraction method based on

speaker identification using wavelet entropy, structures and neural networks; At first, seven Shannon wavelet entropy packets and five structures were extracted from speakers' signals as feature vectors. Unlike traditional speaker recognition methods that derive features from words (or sentences), the proposed method derives features from vowels. Then, these 12 extracted feature coefficients were fed to the neural network as input. Using only 12 feature coefficients, this method achieved 89.16% accuracy.

In 2017, Soleymanpour et al. (Soleymanpour and Marvi 2017) presented a new approach to identify MFCC feature vectors with maximum similarity, which is used to build a speaker recognition model and to define a decision boundary; where MFCC features were extracted from each frame of the speech signal as feature vectors, and then k-means clustering was used to obtain feature vectors with maximum similarity. Experiments were performed using the ELSDSR speech database and neural network was used as a classifier. The results showed that the performance of the speaker identification system has improved in the accuracy criterion.

In 2018, Ali et al. (2018) trained an SVM classifier whose input was selected as a combination of MFCC features with features learned by an unsupervised deep network. The classification accuracy of this method was higher than the method based on non-composite features. The result is 92.6 percent in terms of accuracy in the Urdu language dataset including 10 male and female speakers with 250 samples for each person.

In 2018, Karu and colleagues (Karu and Alumäe 2018) used diarization and i-vector extraction methods with the aim of identifying the speaker in the labeled data at the recording level and trying to label them at the segment level. Using a simple deep neural network with two hidden layers, they trained their model with weak labels. Although this method achieved good accuracy, it is a challenge due to the long time required for training.

In 2018, Hajibabaei and colleagues (Hajibabaei and Dai 2018) used ResNet-20 architecture and logistics marginal cost function. Also, to increase the data in the training and testing stages, they used the techniques of repetition and random slices of reverse time. They managed to achieve an accuracy of 94.6% on the VoxCeleb1 dataset.

In 2020, El-Moneim et al. (El-Moneim et al. 2020), using a short-term memory classifier, designed a system to investigate the effect of MFCC, spectrum, and logarithm-spectrum features under clean and noisy conditions. This system used wavelet denoising and spectral subtraction to increase the detection performance in noisy data. For training, the authors used five female speakers with 100 samples each and achieved 98.7% accuracy with spectrum features and spectrum logarithm on clean data.

In 2020, Rashid Jahangir et al. (Jahangir et al. 2020) proposed a new combination of mel-frequency capstral coefficients and time-based features, which demonstrated the effectiveness of mel-frequency capstral coefficients and time-domain features to improve the accuracy of independent speaker recognition systems. Combines text. The features of Mel-frequency cepstral coefficients and extracted time-based features were given as input to a deep neural network to build a speaker recognition model. The results showed that the mel-frequency cepstral coefficients and time-domain features considered together with the deep neural

network outperform the existing basic mel-frequency cepstral coefficients and time-domain features in the LibriSpeech dataset.
In 2021, Feng Ye et al. (Ye and Yang 2021), proposed a deep neural network model based on a two-dimensional convolutional neural network and gated recurrent unit for speaker recognition. In the design of the network model, the convolution layer was used to extract the acoustic features, which reduces the dimensions of the input in both the time and frequency domains and enables the fast calculation of the gated return layer. Furthermore, the layers of the gated recurrent network were able to learn the acoustic features of a speaker.

Browse the latest works available
Speaker recognition system with self-attention mechanism
In 2019, Neng An and colleagues (An, Thanh, and Liu 2019), inspired by the success of deep neural networks in areas such as speech recognition, speech sentiment analysis, audio event recognition, and image classification, and emphasizing two well-known VGG architectures and ResNet, designed a speaker recognition system. In this system, after applying ResNet and VGG networks, a self-attention layer was added and then the time average integration layer was placed. Similar to other neural network architectures for classification tasks, a softmax layer is used in the final layer. Figure 11-2 shows the complete structure of this network.

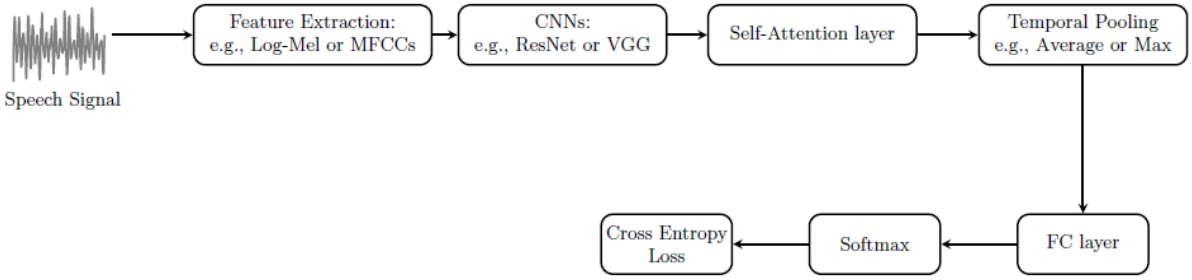


Figure 1. Speaker recognition system with self-attention mechanism proposed by Neng An et al. (An, Thanh, and Liu 2019).

Speaker recognition system inspired by VGG-13 and increasing the number of data

In 2023, Farsiani and his colleagues (Farsiani, Izadkhah, and Lotfi 2022) presented a speaker recognition system inspired by the VGG-13 network, which is shown in Figure 2. In this system, first each audio sample is converted into a log-mel spectrum. This transformation is done for all samples before starting the training process. In each iteration, an increasing number of data is applied to each segment and then the output is fed to the neural network.

The proposed deep neural network similar to VGG-13 consists of 10 convolution layers, except that changes in its architecture, such as reducing the number of filters, have been applied in order to optimize for speaker recognition. The network also has two fully connected layers that classify samples based on their labels in the final layer. This process is repeated until the value of val-loss (i.e. validation error) decreases.

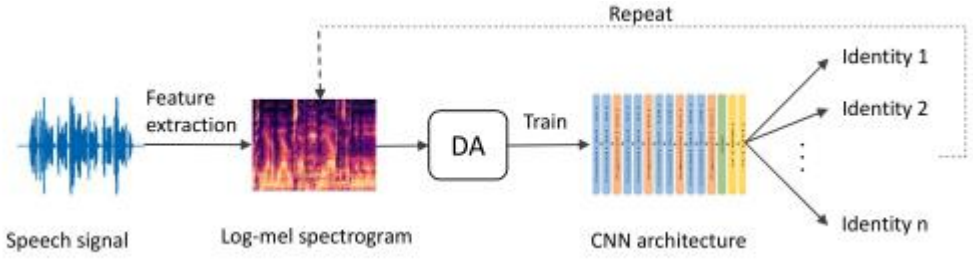


Figure 2. Overview of speaker recognition system inspired by VGG-13 (Farsiani, Izadkhah, and Lotfi .2022)

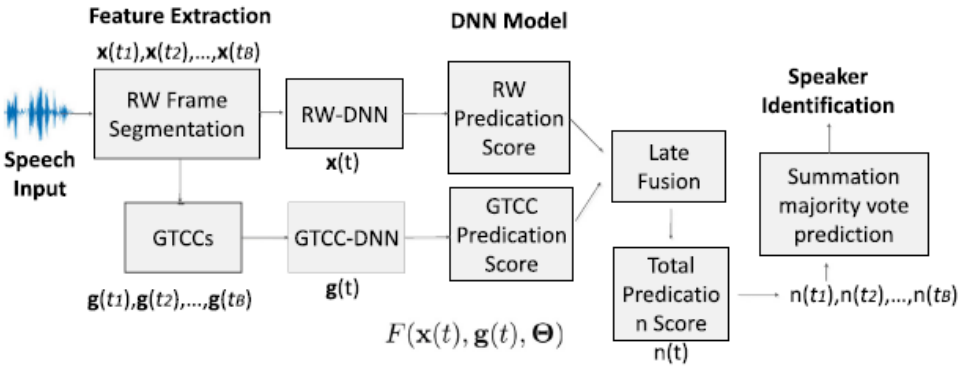
Inspired by the success of VGG in the field of machine vision and speaker recognition, the model architecture was designed based on the VGG-13 version with changes in the number of filters and layers to be more optimal and to match the dimensions of the input image. This architecture consists of 10 convolution layers and kernels of size 3x3 and modified RelU activation function are used in all layers. After each convolution layer, a batch normalization layer is applied, and after both convolution layers, an integration layer with 2x2 steps is added.

After the convolution layers are finished, a global mean pooling layer is applied, followed by a fully connected layer with 4096 neurons. To deal with overfitting, the dropout layer with a rate of 0.5 is placed before and after this layer. Finally, an output fully connected layer with 1251 neurons (equal to the number of speakers) and a softmax cost function is added (Farsiani, Izadkhah, and Lotfi 2022).

Speaker recognition system using time and frequency domain features as input

In 2023, Salvati and colleagues (Salvati, Drioli, and Foresti 2023) assumed that both time-domain and frequency-domain features can be selected as input to a neural network under adverse noise conditions, and thus for a speaker recognition system to be robust. They provided late fusion speaker identification. This system is composed of two independent branches that accept raw short-term waveform and features of Gammaton Capstral coefficients as input, respectively. The model includes a convolutional neural network for feature extraction and a neural network for classification. This research has shown that due to the robustness of the raw waveform features against noise and the features of gammatone capstral coefficients against distortion, the use of common features can significantly improve the performance of the speaker identification system in adverse conditions of noise and distortion compared to the case where only Use a type of feature, improve.

Figure 3 shows the late fusion deep neural network. In this system, the audio signal is divided in the time domain in B short-term frames $\mathbf{x}(t1)$, $\mathbf{x}(t2)$, ..., $\mathbf{x}(tB)$. The frames are used to calculate gammaton capsular coefficients $\mathbf{g}(t1)$, $\mathbf{g}(t2)$, ..., $\mathbf{g}(tB)$ with short-time Fourier transform. Then the raw waveform and gammatone cepstral coefficients are processed by two parallel deep neural networks (RW and GTCC), respectively. Finally, a late fusion layer is used to predict the final class of the speaker. The final prediction score is obtained by summing the late fusion output, $n(t1)$, $n(t2)$,…, $n(tB)$ for all short time frames.



.Figure 3 Overview of the proposed late fusion method (Salvati, Drioli, and Foresti 2023)

The proposed deep neural network model consists of two separate branches, each of which consists of five one-dimensional convolutional layers and three fully connected layers. The activation function used in this network is ReLU. To improve the process of training and normalizing data in categories, after each convolution layer, category normalization is applied. A pooling layer is also used in the RW branch to reduce the dimensionality of the output and extract more robust features.

There is no dimensionality reduction in the convolutional neural network for gammatone capsular coefficients due to the small size of the input. Then, the three fully connected layers compute nonlinear transformations using two dropout layers (with probability 0.5) between them. These fully connected layers have an output size of 512, except for the last layer which has N outputs where N is the number of speakers.

For classification, the cross entropy cost function is used and the Stochastic gradient descent (SGD) method is used to find the model parameters in the optimization problem. Each branch of the deep neural network is trained independently (Salvati, Drioli, and Foresti 2023).

A proposed method inspired by IIR filters for speaker identification

In this method, a new architecture called IIRI-Net was introduced, which is specifically designed for speaker identification. The purpose of this model was to present a network inspired by filters with infinite impulse response (IIR) and to provide a better interpretation of network behavior while maintaining accuracy and efficiency. Due to their special and simple structure, IIR filters can extract key features of audio signals in a more interpretable way.

Challenges and reasons for the development of IIRI-Net

The use of deep neural networks, especially convolutional networks, in the field of speech processing and speaker identification has made significant progress. However, one of the main problems of these models is their low interpretability. Most models act as a black box and it is difficult to explain the reason for their choices and decisions. For this reason, a field called interpretable artificial intelligence (XAI) is trying to develop ways to make complex models and algorithms more understandable and transparent. This issue is especially important in sensitive applications such as identity recognition. The architecture of IIRI-Net was designed

in response to this need and with a focus on improving the interpretability of the basic filters of the network, so that while reducing the complexity of the model, it can provide the necessary information needed to identify the speaker with more transparency.

Characteristics of IIR filters and improve interpretability

In IIRI-Net, instead of using complex filters with more parameters, filters based on IIR, which are usually used in human hearing systems, have been used. These filters allow better interpretation due to their unique properties, such as bandwidth tunability and relative symmetry. IIR filters typically depend on two key parameters: center frequency and bandwidth, which significantly reduces the need for more parameters. In this model, instead of using deep and expensive filters, simpler and interpretable structures of IIR filters are used in the first layer of the network, which works more optimally in terms of the number of parameters, efficiency and accuracy.

Phase correction process in IIRI-Net

In the IIRI-Net architecture, the phase of the filters is also modified to ensure greater interpretability. IIR filters usually have nonlinear phases, which can affect the accuracy of the model; For this reason, in this model, the forward-backward filtering process is used to minimize the phase error. This work causes the phase of the filters to be stabilized in a certain frequency range and the important features of audio signals are recorded more correctly and with more accuracy in the feature extraction stage (Fayyazi and Shekofteh 2023).

Table 4. Summary of said methods

Description	Data set	Results	year	reference number
Using convolutional neural networks with a self-attention mechanism.	voxceleb	Top-1(%) = 90.8 Top-5(%)= 96.5	2019	(An, Thanh, and Liu 2019)
Using a combination of mel-frequency cepstral coefficients and time-based features as input to the system.	librispeech	Accuracy(%)= 89	2020	(Jahangir et al. 2020)
Using a deep neural network based on a two-dimensional convolutional neural network and a gated recurrent unit for speaker recognition.	Aishell-1	Accuracy(%)= 98.96	2021	(Ye and Yang 2021)
	Noise-added Aishell-1	Accuracy(%)= 91.56		
Using a new convolutional neural network for text-independent speaker recognition inspired by the VGG-13 architecture with less parameters but acceptable accuracy.	vocceleb	Top-1(%) = 95 Top-5(%)= 98.1	2022	(Farsiani, Izadkhah, and Lotfi 2022)
The use of meaningful filters was proposed, which was inspired by filters with infinite impulse response. The presented model uses a phase correction process to ensure that phase linearity is satisfied.	TIMIT	CER(%) = 0.8±0.2	2023	(Fayyazi and Shekofteh 2023)
	librispeech	CER(%) = 1.3±0.3		
Application of deep neural network using raw waveform and Capstral gammatone	voxceleb	Accuracy(%)= 80.34	2023	(Salvati, Drioli, and

coefficients as input.				Foresti 2023)

Discussion and Conclusion

Speaker identification, as one of the key issues in speech processing, has experienced many changes in recent decades. In this review article, various speaker identification methods were investigated and the strengths, weaknesses, and applications of each were analyzed. By reviewing the progress of this field, it was found that the use of deep learning techniques, especially convolutional neural networks and networks based on the attention mechanism, have been able to overcome the limitations of classical methods such as Gaussian mixture model and support vector machine to a great extent.

On the other hand, challenges such as speaker identification in noisy conditions, short-term samples and the need for extensive data are still serious obstacles. Investigations showed that hybrid methods, including the use of various features such as MFCC coefficients, Mel-scaled spectrograms, and time-based features, can improve the accuracy and performance of speaker recognition systems.

Finally, this review shows that the future of speaker recognition will move towards using more advanced and flexible networks using new deep learning techniques. Also, designing models with high generalization power and the ability to deal with noise and unbalanced data can provide a path for further progress in this field .

References:

1. Ali, Hazrat, Son N Tran, Emmanouil Benetos, and Artur S d'Avila Garcez. 2018. "Speaker Recognition with Hybrid Features from a Deep Belief Network." *Neural Computing and Applications* 29: 13–19.
2. Almaadeed, Noor, Amar Aggoun, and Abbes Amira. 2015. "Speaker Identification Using Multimodal Neural Networks and Wavelet Analysis." *Iet Biometrics* 4(1): 18–28.
3. An, Nguyen Nang, Nguyen Quang Thanh, and Yanbing Liu. 2019. "Deep CNNs With Self-Attention for Speaker Identification." *IEEE Access* 7(c): 85327–37.
4. Daqrouq, Khaled, and Tarek A Tutunji. 2015. "Speaker Identification Using Vowels Features through a Combined Method of Formants, Wavelets, and Neural Network Classifiers." *Applied Soft Computing* 27: 231–39.
5. El-Moneim, Samia Abd et al. 2020. "Text-Independent Speaker Recognition Using LSTM-RNN and Speech Enhancement." *Multimedia Tools and Applications* 79: 24013–28.
6. Farsiani, Shabnam, Habib Izadkhah, and Shahriar Lotfi. 2022. "An Optimum End-to-End Text-Independent Speaker Identification System Using Convolutional Neural Network." *Computers and Electrical Engineering* 100(January 2021): 107882. https://doi.org/10.1016/j.compeleceng.2022.107882.
7. Fayyazi, Hossein, and Yasser Shekofteh. 2023. "IIRI-Net: An Interpretable Convolutional Front-End Inspired by IIR Filters for Speaker Identification." *Neurocomputing* 558(July): 126767. https://doi.org/10.1016/j.neucom.2023.126767.
8. Hajibabaei, Mahdi, and Dengxin Dai. 2018. "Unified Hypersphere Embedding for Speaker Recognition." *arXiv preprint arXiv:1807.08312*. http://arxiv.org/abs/1807.08312.
9. Jahangir, Rashid et al. 2020. "Text-Independent Speaker Identification through Feature Fusion and Deep Neural Network." *IEEE Access* 8: 32187–202.
10. Karu, Martin, and Tanel Alumäe. 2018. "Weakly Supervised Training of Speaker Identification Models." *Speaker and Language Recognition Workshop, ODYSSEY 2018*: 24–30.
11. Salvati, Daniele, Carlo Drioli, and Gian Luca Foresti. 2023. "A Late Fusion Deep Neural Network for Robust Speaker Identification Using Raw Waveforms and Gammatone Cepstral Coefficients." *Expert Systems with Applications* 222(February): 119750. https://doi.org/10.1016/j.eswa.2023.119750.
12. Soleymanpour, Mohammad, and Hossein Marvi. 2017. "Text-Independent Speaker Identification Based on Selection of the Most Similar Feature Vectors." *International Journal of Speech Technology* 20: 99–108.
13. Ye, Feng, and Jun Yang. 2021. "A Deep Neural Network Model for Speaker Identification." *Applied Sciences (Switzerland)* 11(8).