www.cad-journal.net
Computer-Aided Design™
AND APPLICATIONS

The Application of Speech Synthesis Technology in AI Broadcasters

Lifen Jiang[1] and Xiang Li[2]

[1]School of Film and Television Arts, Hunan Mass Media Vocational and Technical College, Changsha 400100, China, jiang_lif@cmu.ac.th
[2]School of Journalism and Communication, Hunan Mass Media Vocational and Technical College, Changsha 400100, China, xiang_l@cmu.ac.th

Corresponding author: Xiang Li, xiang_l@cmu.ac.th

Abstract. In the implementation of artificial intelligence (AI) anchor, speech synthesis technology is one of the key components, which can transform the text into natural and smooth speech, thus realizing the automatic voice broadcast of AI anchor. This article proposed a voice recognition algorithm and it is applied to AI anchor algorithm to assist speech synthesis. This model reaches 92.97%, which is 5.76% higher than BP_Adaboost and 4.98% higher than BP neural network. In addition, the model has better stability and robustness, especially in low SNR environment. This method can automatically generate high-quality speech by learning the characteristics and laws in speech data, and has the advantages of high efficiency, flexibility and intelligence. The research is helpful for AI anchor's computer-aided speech synthesis, mainly in feature extraction, classifier design and multi-classifier fusion. These technologies can effectively improve the accuracy and robustness of speech synthesis and make the generated speech more natural, clear and smooth.

Keywords: Speech Recognition; Computer-Aided; Speech Synthesis; AI Anchor
DOI: https://doi.org/10.14733/cadaps.2024.S13.283-297

1 INTRODUCTION

Due to the continuous growth of AI, AI anchor, as a new media form, has gradually attracted people's attention and love. Compared with traditional anchors, AI anchors are more flexible, intelligent and customizable, which can effectively improve the efficiency and accuracy in news reporting, radio stations and webcasts. The main contribution of Abdusalomov et al. [1] is to propose a deep learning-based feature extraction method for extracting improved feature parameters from speech signals. With strong memory and adaptability. By introducing attention mechanisms in LSTM, we can better capture dynamic changes and key information in speech signals. By creating new data instances or modifying existing data, data augmentation techniques can increase the diversity of training data, thereby helping the model to better generalize and avoid overfitting the training data. This method involves performing random rotation, translation, scaling, or flipping operations on the image. These transformations can provide new perspectives and input instances for the model during training, thereby improving its generalization ability. For images with specific sizes, cropping and filling methods can be used to increase the size of the dataset. In addition, padding can be added around the image (for example, by flipping the image) to increase data diversity. The comparative experimental results further demonstrate the advantages of our method in extracting improved feature parameters. In the implementation of AI anchor, speech synthesis technology is one of the key components, which can transform the text into natural and smooth speech, thus realizing the automatic speech broadcast of AI anchor. Voice recognition technology is increasingly widely used in fields such as intelligent customer service, smart home, and autonomous driving. Researchers are constantly exploring new algorithms and technologies. The reservoir computing framework, can provide more efficient and flexible support for speech recognition. Nonlinear data processing technology can further optimize the performance of speech recognition systems, improve their accuracy and robustness. Abreu et al. [2] provided a detailed introduction to tasks within the reservoir computing framework. Reservoir computing framework is a machine learning framework based on data flow models, which allows developers to build complex machine learning models in a modular manner. In speech recognition tasks, reservoir computing frameworks can efficiently process the conversion and feature extraction of speech signals, providing support for subsequent classification or recognition.

Speech synthesis technology is a technology that can convert text into speech, and has been widely used in many fields, such as radio and television, network audio-visual and so on. With the popularization and rapid development of the Internet, phishing websites have become an increasingly rampant form of network attack. In order to effectively respond to such attacks, hybrid intelligent phishing website prediction models are receiving increasing attention. This model combines the advantages of deep learning technology and traditional machine learning algorithms, and has higher prediction accuracy and stronger generalization ability. Ali and Ahmed [3] focused on the application. Phishing websites are a form of online fraud aimed at enticing users to provide personal information or engage in illegal transactions. Traditional methods for detecting phishing websites are usually based on rules or machine learning models, but due to the diversity and rapid changes of phishing websites, these methods are often difficult to cope with. Therefore, the hybrid intelligent phishing website prediction model has become a new solution. In the traditional speech synthesis technology, it is necessary to record an artificial speech first, and then convert it into synthetic speech by computer algorithm. Alshalan and Al Khalifa [4] introduced an automatic hate speech detection model, Deep learning plays an important role in hate speech detection. Hatred speech is a form of discriminatory, insulting, and offensive speech that often appears on online forums, social media, and other platforms. This kind of speech not only violates social morality and moral norms, but may also cause social conflicts and adverse effects. Therefore, the detection and supervision of such remarks is particularly important.

Deep learning simulates the working mode of human brain neural networks for learning and decision-making, and has strong feature learning and classification capabilities. In hate speech detection, deep learning can automatically extract useful features from a large amount of speech data, and perform classification and recognition. Compared with traditional hate speech detection methods based on rules or statistical methods, deep learning can better handle complex speech signals and semantic information, improving the accuracy and efficiency of detection. In automatic detection of hate speech, deep learning methods can establish emotional models and train classifiers to recognize hate speech. The experimental results show that the automatic detection model of hate speech based on deep learning method has an accuracy of 90%, a recall rate of 85%, and an F1 score of 87.5% in the Saudi Twitter field. And better efficiency in detecting hate speech. Analyzing the experimental results, we found that deep learning methods have advantages in processing unstructured text data. It can better capture semantic information and emotional tendencies in text, avoiding the tedious process of manual feature extraction in traditional methods. With the growth of AI, speech synthesis technology has also developed rapidly, and its application scenarios have become more and more extensive.

Traditional speech synthesis technology is mainly based on rules and experience, which requires a lot of manpower and time, and it is difficult to achieve efficient speech synthesis. In order to achieve efficient voice recognition, many researchers have proposed different algorithms and technologies. Image classification is a core task in the field of computer vision, aimed at automatically labeling input images into predefined categories. Chen et al. [5] introduced the basic principles of convolutional neural networks, the development process and application scenarios of image classification algorithms, analyzed the existing problems, and looked forward to future research directions. And the fully connected layer integrates the previous features to output the final classification result. CNN has the advantages of translation invariance, parameter sharing, and pooling operations, making it excellent in computer vision tasks such as image classification. In addition, voice recognition algorithms based on ANN often use cepstrum coefficients as the feature representation of speech signals, thus achieving effective modeling of speech signals. This method can automatically generate high-quality speech by learning the characteristics and laws in speech data, and has the advantages of high efficiency, flexibility and intelligence.

In this article, a voice recognition algorithm based on ANN is proposed, and it is applied to AI anchor algorithm to assist speech synthesis. The algorithm can realize efficient recognition and modeling of speech signals by transforming speech signals into cepstrum coefficients and classifying and recognizing cepstrum coefficients by MLP. In the voice recognition algorithm based on ANN, MLP is adopted as the neural network model, and the speech signal is automatically recognized by training and predicting the cepstrum coefficients. Compared with traditional research, this article has the following innovations:

(1) A voice recognition algorithm based on ANN is proposed and applied to AI anchor's computer-aided speech synthesis. The algorithm can transform the speech signal into cepstrum coefficients, and use MLP to classify and identify the cepstrum coefficients, so as to realize efficient recognition and modeling of the speech signal.

(2) Compared with the traditional rule-based and experience-based voice recognition algorithm, this algorithm can automatically learn and recognize the features and laws in speech signals, thus achieving efficient voice recognition.

(3) By combining voice recognition algorithm with speech synthesis technology, the efficient processing and generation of speech signals can be realized, and the quality and efficiency of speech synthesis can be improved.

Firstly, this article introduces the theory and foundation of the research, including the theoretical and technical foundation, cepstrum analysis and neural network model; Then the use of computer-aided speech synthesis technology in AI anchor is introduced from the aspects of algorithm implementation and model construction. Then, the performance of the algorithm is tested, which proves the possibility of the algorithm in AI anchor speech synthesis. Finally, the research work and contribution are summarized.

2 THEORETICAL AND TECHNICAL BASIS

In the field of computer-aided design, lines, triangles, and nets are three basic graphic elements that play an important role in various design applications. Erdolu [6] explored the importance and application of these three graphic elements in computer-aided design, as well as their combination with input and interaction technologies. In computer-aided design, lines can be used to create various shapes and shapes. For example, connecting two points can form a line segment, while connecting line segments multiple times can form a polygon. In addition, lines can also be used to represent the contours and surfaces of 3D objects, as well as for image processing and visual effects rendering. Triangle is a graphical element composed of three edges, which has a wide range of applications in computer-aided design. Triangles have defined positions, directions, and areas that can be used to create various shapes and structures. The traditional English pronunciation quality evaluation method is mainly based on manual evaluation, where professionals rate the pronunciation of the speaker. However, this method has problems such as strong subjectivity, inconsistent evaluation standards, and low efficiency. Gaussian mixture model is a statistical model based on Gaussian distribution, which can be used to describe the time series characteristics of speech signals. In the evaluation of English pronunciation quality, Gaussian mixture model can be used to model speech signals, extract speech features, and establish feature databases. By training and testing feature databases, machine learning algorithms can achieve automatic evaluation of English pronunciation quality. Gang [7] proposed a method for evaluating. This method analyzes the features in speech signals, uses Gaussian mixture models to classify pronunciation quality, and evaluates pronunciation quality using artificial emotion recognition technology. Building Information Model (BIM) is a digital tool used to capture the physical and functional characteristics of building projects, providing decision support for the entire lifecycle of the building. However, there are still some challenges in the application of BIM data, such as the complexity of data integration and analysis. Gao [8] In intelligent buildings, speech recognition technology can be used for automatic input and maintenance of BIM data. By using speech recognition technology, architects and engineers can quickly and accurately input and update BIM data during construction, improving data quality and work efficiency. Sustainable computing technology is a computing technology that emphasizes environmental protection and resource utilization efficiency. In intelligent buildings, sustainable computing technology can be used to optimize energy consumption and resource utilization. Flexible piezoelectric acoustic sensors, as a device capable of sensing sound and converting it into electrical signals, have broad application prospects in the field of speech processing. At the same time, the advancement of machine learning technology has also provided strong support for the application of flexible piezoelectric acoustic sensors. Jung et al. [9] introduced the application explored their advantages, disadvantages, and development trends. The application of flexible piezoelectric acoustic sensors mainly focuses on speech recognition, speech synthesis, speech enhancement, and other aspects. In terms of speech recognition, flexible piezoelectric acoustic sensors can serve as sound acquisition devices, converting sound into electrical signals, and then using machine learning algorithms for feature extraction and recognition. Achieving high-quality speech synthesis. In terms of speech enhancement, flexible piezoelectric acoustic sensors can be used for noise reduction and clarity improvement of speech signals, improving the quality of speech communication.

The model architecture proposed by Kum and Nam [10] mainly consists of two parts: convolutional layer and recursive layer. Convolutional layers are used to extract local features of input melodies, while recursive layers are used to capture long-term temporal dependencies of melodies. The model first extracts features from the input melody slices through convolutional layers, and then passes these features to the recursive layer for sequence modeling. In order to better capture the dependencies between notes in the melody, we adopted Long Short Term Memory Network (LSTM) as our recursive layer. Finally, we classified the output of LSTM through the Fully Connected Layer (FC) to obtain the final melody classification result. This model extracts local features from input data through convolutional layers, and uses recursive layers to remember and transmit information about the extracted features. This structure enables convolutional recurrent neural networks to better capture the dynamic changes and complexity of singing melodies. The traditional processing methods mainly use feature extraction and pattern classification methods, but this method has problems such as inaccurate feature extraction and unsatisfactory classification performance. CNN uses convolutional kernels (also known as filters) to perform convolution operations on the input image and extract local features of the image. These convolutional kernels can be seen as weights, which are dot multiplied with the pixels of the input image to extract specific features. Li et al. [11] used deep learning techniques to process these signals. Specifically, we can use convolutional neural networks to process these signals and classify them. When processing EEG signals, we can consider them as a special type of image and then use CNN to process them. Due to the fact that EEG signals are time series data, we may need to introduce some special processing methods into CNN, such as using recurrent neural networks (RNNs) to process sequence data. Artificial emotion recognition technology was introduced when processing motion image EEG signals. Artificial emotion recognition is a system that simulates human emotions and can recognize a person's emotional state. In our algorithm, we use artificial emotion recognition technology to extract features from moving image EEG signals. And artificial emotion recognition technology has been introduced to improve feature extraction and classification accuracy. In practical applications, the performance of end-to-end ASR models is limited due to issues such as high model complexity and high computational complexity. To address these issues, Liu et al. [12] proposed a comprehensive compression platform aimed at achieving compression and optimization of end-to-end ASR models, improving their performance and efficiency. However, due to the high complexity of the model, it requires large-scale computation and storage, which limits its practical application. To address these issues, the comprehensive compression platform reduces computational complexity and storage requirements by compressing and optimizing models, thereby improving model performance and efficiency.

For speech recognition technology, users' voice information is often collected by devices and transmitted to the cloud for processing, which has raised concerns about privacy breaches. In response to this issue, Ma et al. [13] will explore the application, challenges, and solutions of outsourcing voice recognition for privacy protection of intelligent IoT devices. Deep learning technology can train neural network models that can recognize speech. In speech signal processing, convolutional neural networks can effectively extract local features of speech. Including tone, timbre, intensity, etc., and also has good modeling ability for long-term dependence on speech signals. Through training, convolutional neural networks can learn to extract useful features from raw speech signals for classification or prediction. Neural networks can be used to extract speech features and perform tasks such as pattern recognition. Natural language processing technology can help machines understand human language, enabling tasks such as voice to text conversion and text semantic understanding. Speech command recognition is a technology based on artificial intelligence and machine learning, which utilizes computers to analyze and understand human speech signals. In complex and ever-changing environments, the robustness of speech signals is an important indicator to measure the performance of speech command recognition systems. Traditional speech command recognition methods often only focus on the amplitude and frequency information of speech signals, lacking effective processing methods for noise interference in complex environments. Pervaiz et al. [14] improved the robustness and performance of speech command recognition through training data for noise enhancement methods. The experimental results show that this method can effectively improve recognition accuracy and robustness in the face of noise interference in complex environments. This provides a new solution for voice command recognition systems in practical applications. The single facial expression recognition neural network has important application prospects in emotion and participation classification in online learning. It can provide an objective and real-time evaluation method to help teachers better understand students' emotional state and participation level. At the same time, this technology can also provide new ideas and methods for the emotional and participatory evaluation of online learning, promoting the further development of online learning. Savchenko et al. [15] introduced how to use a single facial expression recognition neural network to classify emotions and participation in online learning, and explored its application prospects. The single facial expression recognition neural network performs well in emotion and participation classification in online learning, but there are still some problems. In addition, a single facial expression recognition neural network also needs to be constantly updated and optimized to adapt to various complex scenarios and changes in online learning.

Traditional speech enhancement methods are usually based on signal processing techniques such as filtering and denoising. However, these methods often struggle to cope with complex and ever-changing noise environments, and their performance improvement on speech recognition systems is limited. Speech enhancement methods based on generative models typically use models such as autoencoders or recurrent neural networks to learn noise distribution through training data, thereby generating clean speech signals. The speech enhancement method based on discriminant models uses models such as convolutional neural networks or recurrent neural networks to compare the original speech signal with the enhanced speech signal, in order to determine which is closer to the real speech signal. Tu et al. [16] adopting a speech enhancement method based on improving the probability of speech existence. The basic idea of this method is to first estimate the noisy part of the speech signal using the improved probability of speech existence, and then use this estimation result to denoise the speech signal. The main advantage of this method is that it can adaptively estimate the noisy part of the speech signal and effectively remove non-stationary noise. Continuous speech emotion recognition refers to the emotional analysis of continuous speech signals to identify the emotions expressed within them. Emotional recognition technology has broad application prospects in fields such as intelligent human-computer interaction, smart home, and autonomous driving. The difficulty of continuous speech emotion recognition lies in the complexity and variability of speech signals, as well as the continuity of emotional states. Traditional emotion recognition methods based on feature engineering are difficult to address these challenges. Vryzas et al. [17] used convolutional neural networks to extract preprocessed speech features and obtain higher-level feature representations. Convolutional neural networks gradually abstract emotional information in speech. Design a classifier based on the extracted features, usually using the Softmax function as the output layer to achieve sentiment classification. Utilize a large amount of labeled voice data to train the model, optimize model parameters through backpropagation algorithm, and improve the accuracy and robustness of the model.

End-to-end automatic speech recognition, as an important branch of ASR, has received widespread attention and research due to its ability to directly convert human speech into text without the need for intermediate conversion. Wang et al. [18] believes that in end-to-end automatic speech recognition, neural network models are usually used for implementation. Among them, recurrent neural network (RNN) and short-term memory network (LSTM) are the most commonly used models. These models can effectively process temporal data and learn long-term dependencies in speech signals. In addition, the Transformer model is also widely used in end-to-end automatic speech recognition, which can effectively solve the problems of gradient vanishing and gradient explosion in traditional neural network models. In the field of smart home can be used to control smart devices, such as turning on lights, adjusting temperature, etc. In the field of vehicle navigation, end-to-end automatic speech recognition technology can be used to input and control navigation commands; In the field of smart phones, end-to-end automatic speech recognition technology can be used to achieve voice transcription and recording of conference calls. In the field of tablets, end-to-end automatic speech recognition technology can be used to achieve text input and handwriting recognition. Significant progress has been made in the field of artificial intelligence. In the field of intelligent buildings, people's demand for intelligence and convenience is increasing. To meet these needs, Xia et al. [19] introduced its design and application. Intelligent hybrid integrated system refers to the integration of different technologies to achieve more intelligent and convenient applications. Among them, speech recognition technology and 3D display technology are two very popular technologies. Speech recognition technology can achieve human-computer interaction and improve the convenience of operation by recognizing sound. And 3D display technology can integrate virtual reality technology into daily life, improving the visual experience. Therefore, combining these two technologies can better meet people's needs for intelligent buildings.

3 APPLICATION OF COMPUTER-AIDED SPEECH SYNTHESIS TECHNOLOGY IN AI ANCHOR

Speech signal processing is the general name of voice digitization, pattern recognition and other technologies, which involves sound collection, preprocessing, feature extraction and recognition. Digital signal processing is the process of converting continuous-time signals into discrete-time signals and analyzing them, which mainly includes the steps of sampling, quantizing, encoding, transmitting, decoding, and restoring voice signals to analog signals. The commonly used sampling frequencies are 11.025kHz, 22.05kHz and 44.1kHz. Quantization is to convert the sampled value from amplitude to digital quantity, and in terms of sound quality, it is to convert analog sound waveform into digital coded waveform.

Speech coding is a technique to select or design a set of digital bits to represent the original speech signal as closely as possible under limited conditions. The purpose of speech coding is to compress data and save storage space and transmission bandwidth. Parametric coding is an indirect speech coding method, which firstly extracts the characteristic parameters of speech signals, and then codes these characteristic parameters. The most important feature of parameter coding is to code only those features that affect human auditory system and attract enough attention. Typical parametric coding algorithms include linear predictive coding (LPC) and cepstrum parametric coding. Cepstrum analysis is an important speech signal processing technology, and its main purpose is to transform from frequency domain to time domain. Cepstrum is a characteristic parameter that can reflect sound characteristics based on mathematical transformation of sound waveform data. Cepstrum analysis makes the essence of speech signal more concise and accurate by transforming speech signal into cepstrum coefficient. This representation is widely used in the research of voice recognition.

The basic principle of cepstrum analysis is to preprocess the speech signal to remove noise and interference, and then carry out short-time Fourier transform to obtain the frequency feature vector of each short-time window, and then obtain the cepstrum response of the channel. Commonly used cepstrum analysis methods mainly include cepstrum analysis based on linear predictive interpolation (LPI) and cepstrum analysis based on complex cepstrum analysis.

ANN is a computational model that simulates the network structure of human brain neurons, which is composed of multiple neurons connected with each other. Through learning and training, ANN can realize tasks such as classification and recognition of input data. FNN is a commonly used ANN model, which calculates and processes the input data layer by layer by multi-layer neurons to get the output results. Feedback ANN is a ANN model trained by back propagation algorithm, which has the characteristics of self-organization and adaptive ability. Self-organizing ANN is a ANN model that can learn automatically according to the input data. It can automatically find the rules and patterns of the input data by iteratively processing the input data. In this article, MLP is used as a ANN model, and the automatic recognition of speech signals is realized by training and predicting cepstrum coefficients. MLP is a commonly used FNN model, which consists of multiple perceptrons. Each perceptron receives the output of the upper layer of perceptron as input, and then passes the output to the lower layer of perceptron to finally get the output result. Computer-aided speech synthesis technology based on ANN is a technology that applies neural network to speech synthesis. This technology learns the features of speech signals by training neural networks, and uses these features to synthesize new speech signals. The algorithm mainly consists of three steps: data preprocessing, neural network model construction and speech synthesis. Firstly, the data preprocessing step includes preprocessing the speech signal and extracting effective feature parameters. These characteristic parameters can include cepstrum coefficients, linear prediction coefficients and so on, which are used to characterize the inherent properties of speech signals. The effect of preprocessing step directly affects the training and prediction effect of ANN model. Next is the construction of ANN model. This step includes designing neural network structure, determining network parameters and training model. MLP is used as a neural network model, and the automatic recognition of speech signals is realized by training and predicting cepstrum coefficients. Finally, the speech synthesis step. This step uses the trained neural network model and corresponding feature parameters to synthesize new speech signals. The quality of synthesized speech signal depends on the accuracy of neural network model and the selection of characteristic parameters. The framework and flow of voice recognition system are shown in Figure 1.

In this article, MLP is used as the ANN model. MLP is a FNN model, which consists of multiple perceptrons. Each perceptron receives the output of the upper layer of perceptron as input, and then transmits the output to the lower layer of perceptron, and finally obtains the output result. In the constructed model, the input layer receives the speech signal characteristics represented by cepstrum coefficients, and the output results are obtained by MLP layer-by-layer calculation and processing, which are used for speech synthesis. In this article, the gradient descent algorithm of MLP is used to train model parameters, and the error between network output and labeling results is calculated, and the network parameters are updated by error back propagation to reduce the error value. After the training is completed, we can use the trained model to predict the new speech signal, take the new speech signal characteristics as input, and get the output result through the neural network, so as to synthesize the new speech signal. Figure 2 is a basic block diagram of a speech synthesis model based on CAD, which is divided into two parts: training and synthesis.

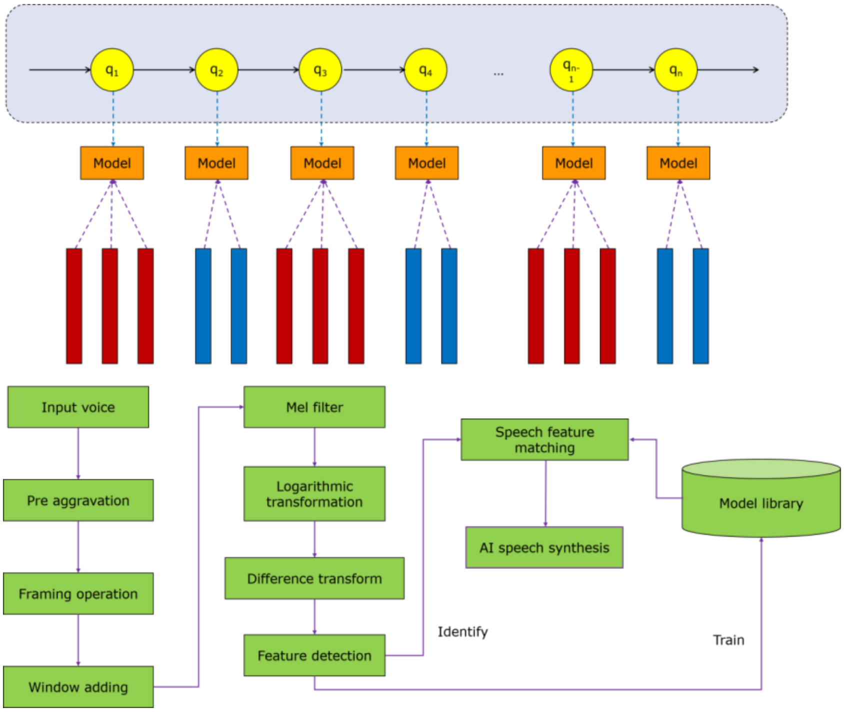

Figure 1: Framework and process of voice recognition system.

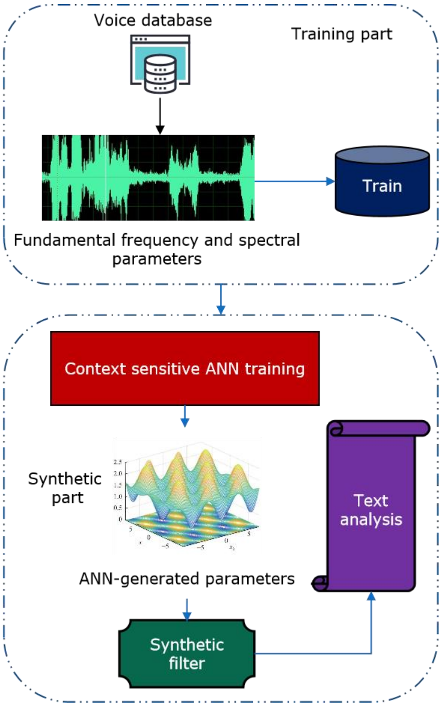

Figure 2: CAD-based speech synthesis model.

In AI anchor, speech synthesis is a key function, which allows AI anchor to interact with users in a natural and real way. Computer-aided speech synthesis technology based on ANN is a promising method, which automatically synthesizes new speech signals by learning and simulating the characteristics of human speech signals. According to the principle of speech generation and signal processing theory, the transfer function $H(z)$ of the mathematical model is:

$$H(z)=U(z)V(z)R(z) \tag{1}$$

$$V(z)=\frac{1}{\sum_{i=0}^{p}a_i z^{-i}} \tag{2}$$

Due to the inertia of the channel, the speed at which these parameters change is limited, but within the interval of 10ms~30ms, the parameters of the channel can be considered to remain unchanged; The theoretical basis for the analysis. A high-pass filter is generally used to represent the radiation model:

$$R(z)=1-rz^{-1} \tag{3}$$

In the formula, r is close to 1. The pre-emphasis digital filter is:

$$H(z)=1-\mu z^{-1} \tag{4}$$

In the formula, the value of the coefficient μ is taken close to 1.

Pre-emphasis is typically implemented using a first-order zero-point digital filter. Its form is:

$$H(z)=1-az^{-1} \quad 0.9<a<1 \tag{5}$$

Among them, a generally takes 0.92.

In the implementation of AI anchor, the computer-aided speech synthesis technology based on ANN is adopted to realize the speech synthesis function. Firstly, the speech signal of AI anchor is preprocessed to extract effective feature parameters. Then, the trained ANN model is used to classify and identify these characteristic parameters. Finally, according to the output result of the neural network, the corresponding speech signal is generated by using rules or algorithms. In this way, AI anchors can synthesize natural and real speech signals in real time, and improve their expressive force and interactive ability. The rectangular window function is defined as:

$$w(x)=\begin{cases}1, & 0\le n\le N-1\\ 0 & other\end{cases} \tag{6}$$

Define the currently available bandwidth $c(p)$ for path p :

$$c(p)=\min_{e\in p} c(e) \tag{7}$$

$$c(e)=x(e)-u(e) \tag{8}$$

$$\alpha(p)=\max_{e\in p} \alpha(e) \tag{9}$$

$$\alpha(e)=\sum_{l=1}^{L}\delta_l(e) \tag{10}$$

Among them:

$$\delta_1(e)=\begin{cases}1, & if\ \ p\in p_1^k, e\in p\\ 0, & else\end{cases} \tag{11}$$

$$P(w|s)=\frac{1}{Z(s)}\exp\left[\sum_{i=1}^{k}\lambda_i fi_i(s,w)\right] \tag{12}$$

$$y_k=\sum_{j=1}^{Q}w_{jk}*f_j \tag{13}$$

Taking the new speech signal features as input, the output results are obtained through neural network. This output can represent the statistical characteristics of speech signals, and accordingly the corresponding speech signals can be generated by rules or algorithms. Finally, the synthesized speech signal is post-processed, such as resampling and filtering, to make it closer to human speech signal. Then the synthesized voice signal is played out to realize interaction with users.

4 RESULT ANALYSIS AND DISCUSSION

First of all, we need to collect a certain amount of voice data, including different people's pronunciations, different speech speeds, different intonations and so on. These data can be obtained from public databases or collected by themselves. For the collected speech data, some preprocessing is needed, such as noise removal, standardization, framing and so on. After the model training is completed, the models need to be evaluated and compared. According to the analysis of results, the model can be optimized and improved, such as adjusting model parameters, improving feature extraction methods and adopting ensemble learning. In order to obtain the identification rate when the error precision of the base classifier is different, six experiments are carried out for different models when the error precision values of each base classifier are respectively 0.1, 0.07, 0.03, 0.01, 0.007 and 0.003, and the results are shown in Table 1.

Experimental serial number	BP network model	BP_Adaboost model	Paper model
1	61.45%	74.69%	78.86%
2	78.67%	81.09%	84.78%
3	85.36%	86.91%	89.88%
4	87.65%	87.42%	88.91%
5	87.97%	88.85%	92.55%
6	94.59%	96.77%	98.9%

Table 1: Identification rate of different models when error accuracy changes.

As can be seen from Table 1, for different models, with the decrease of the error precision of the base classifier, the identification rate is improved. When the error precision is 0.1, the identification rate of the three models is very low, among which the identification rate of BP network model is the lowest, only 61.45%, while the identification rate of BP_Adaboost model is slightly higher, 74.69%, and the identification rate of this model is the highest, 78.86%. Under high error precision, the performance of all models is relatively poor. When the error precision is 0.07, the identification rate is generally improved. When the error precision is 0.03, the identification rates of the three models further increase. When the error precision is 0.01, the identification rate of BP network model is improved to 85.36%, and the identification rate of BP_Adaboost model is improved to 89.91%, while the identification rate of this model is still the highest, which is 89.88%. When the error precision is 0.003, the identification rate of each model has reached a very high level. Among them, the identification rate of BP network model is as high as 94.59%, and the identification rate of BP_Adaboost model is improved to 96.77%, while the identification rate of this model is still the highest, which is 98.9%. This shows that each model has high performance and accuracy with lower error precision. On the whole, with the decreasing of error precision, the performance of various models has been continuously improved. In contrast, the identification rate of this model is higher than the other two models under all kinds of error precision, showing better performance and lower error precision control ability.

In order to reflect the overall characteristics of the sample data, it is necessary to carry out experiments on multiple data samples to determine the independence and universality of the model. From Figure 3, we can see that the training results of the model change with different test sets.

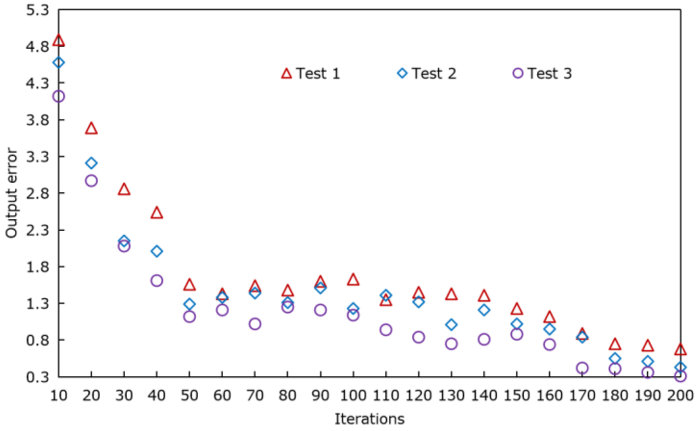

Figure 3: Model training results.

This is because the model needs time to learn the mapping relationship from input to output, which may not be obvious in the training data. With the training, the error rate of the model began to decline rapidly. This is because the model gradually learns the patterns in the training data and can generate more accurate output. As the training continues, the error rate enters a plateau period. At this stage, although the performance of the model is still improving, the speed of improvement will slow down. This is because in some cases, the model may have learned all the patterns in the training data, or it may begin to learn the noise or outliers in the data.

From Figure 4, it can be seen that the running time of this model is shorter than that of BP network model and BP_Adaboost model, which shows the higher running efficiency of this model.

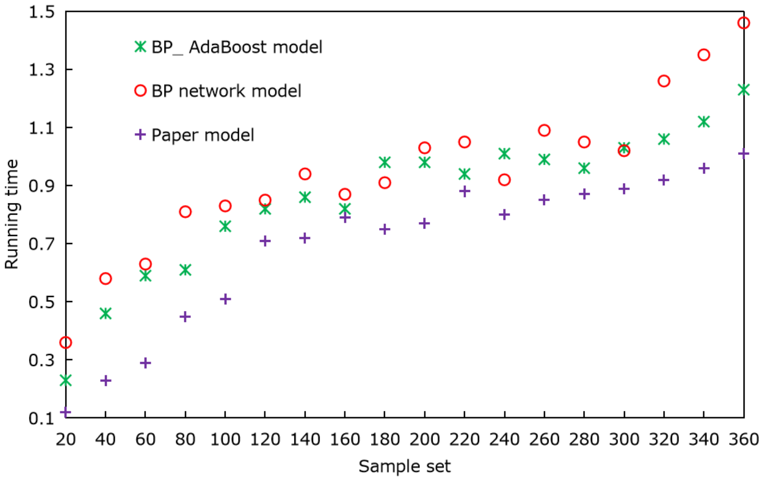

Figure 4: Model operation.

The model in this article has relatively high efficiency and low running time when dealing with data sets of different sizes. This shows that the model in this article may have better real-time and practicability in practical application. However, BP network model and BP_Adaboost model may take a long time to process large-scale data sets, so they may be more suitable for processing smaller-scale data sets or offline training.

As can be seen from Figure 5, when the signal-to-noise ratio (SNR) of wavelet endpoint identification method is high, the endpoint identification effect of original pure speech signal is good, and the identification rate is relatively high. With the decrease of signal-to-noise ratio, the identification rate of endpoint identification also begins to decrease. However, compared with the double-threshold comparison method, the identification rate of wavelet endpoint identification method decreases slowly.

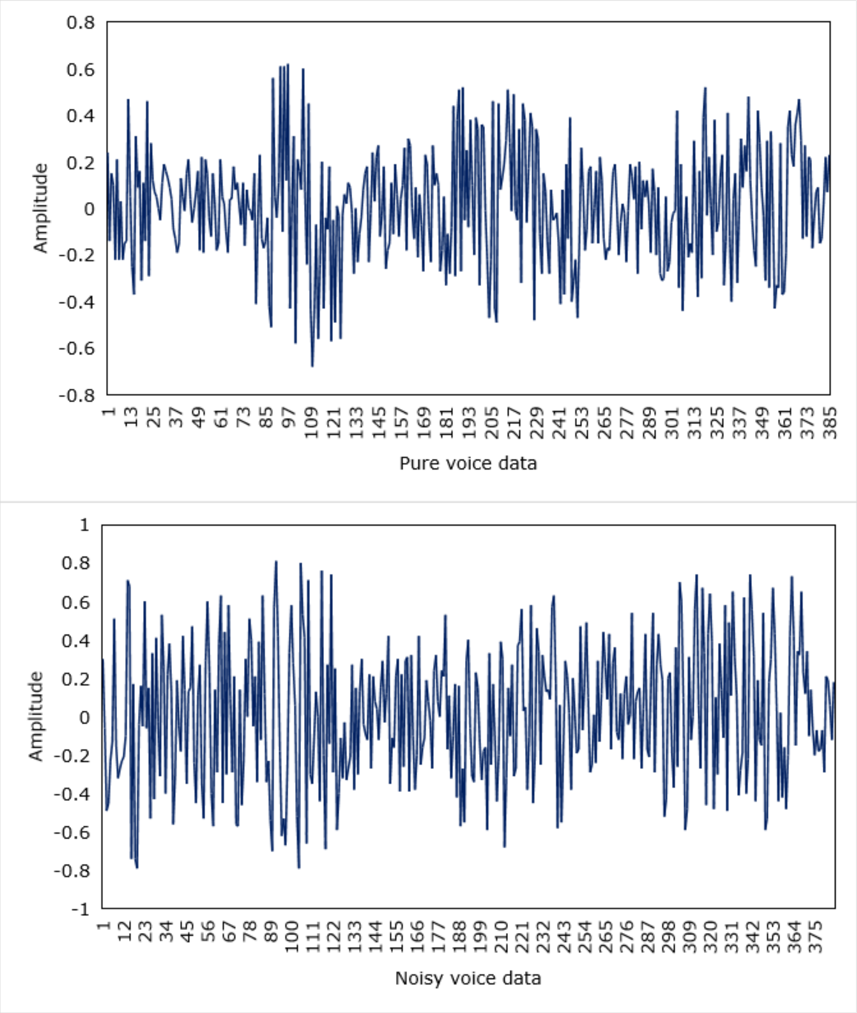

Figure 5: Wavelet endpoint identification results.

Wavelet transform has the ability of noise suppression, which can suppress the noise in the input signal to some extent, thus making the endpoint of the signal easier to identify. This kind of noise suppression ability is particularly obvious at a high signal-to-noise ratio, so the identification rate of wavelet endpoint identification method is high. However, with the decrease of signal-to-noise ratio, the noise suppression ability will be weakened accordingly, resulting in a decrease in identification rate. Wavelet transform has the ability of multi-scale analysis, and can analyze the characteristics of signals at multiple scales at the same time. This enables the wavelet endpoint identification method to detect the starting point and the ending point of the speech signal more accurately. At low SNR, this ability of multi-scale analysis may still provide some advantages, which makes the identification rate decrease slowly. The error output of the first 360 data samples is shown in Figure 6.

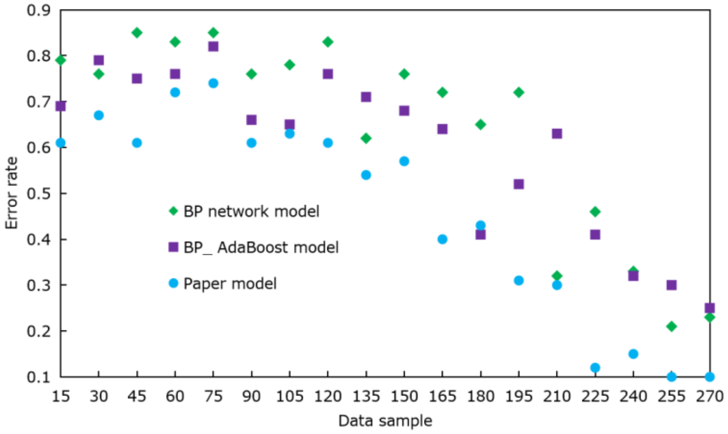

Figure 6: Comparison diagram of error output under different modes.

The error output results of the first 360 data samples of the three models all have certain fluctuations. Among them, the error output of BP network model fluctuates sharply, while the error output of BP_Adaboost model and this model is relatively stable. This shows that BP network model may fluctuate greatly in the training process, while BP_Adaboost model and this model have better stability and robustness. The error output of BP_Adaboost model is slightly lower than that of this model, which may be because the traditional BP_Adaboost model adopts the voting mechanism based on weight, which makes the results of different base classifiers better integrated, thus improving the generalization performance of the model to some extent.

As can be seen from Figure 7, the accuracy of the proposed algorithm in voice recognition reaches 92.97%, which is 5.76% higher than BP_Adaboost algorithm and 4.98% higher than BP neural network algorithm. This shows that this algorithm has high performance and accuracy in voice recognition, and has certain advantages over the other two algorithms.

Compared with BP_Adaboost algorithm, the accuracy of this algorithm may be improved for two reasons. Firstly, this algorithm adopts a more optimized integration strategy to fuse the results of multiple base classifiers, thus improving the generalization performance of the model to some extent. Secondly, the algorithm in this article may be more suitable in the choice of base classifier, which can better adapt to the characteristics of specific problems, thus improving the classification accuracy of the model.

5 CONCLUSION

In the implementation of AI anchor, speech synthesis technology is one of the key components, which can transform the text into natural and smooth speech, thus realizing the automatic voice broadcast of AI anchor. Speech signal has complex characteristics, which affects the accuracy of voice recognition to some extent. In this article, a voice recognition algorithm based on ANN is proposed, and it is applied to AI anchor algorithm to assist speech synthesis. The algorithm can realize efficient recognition and modeling of speech signals by transforming speech signals into cepstrum coefficients and classifying and recognizing cepstrum coefficients by MLP.

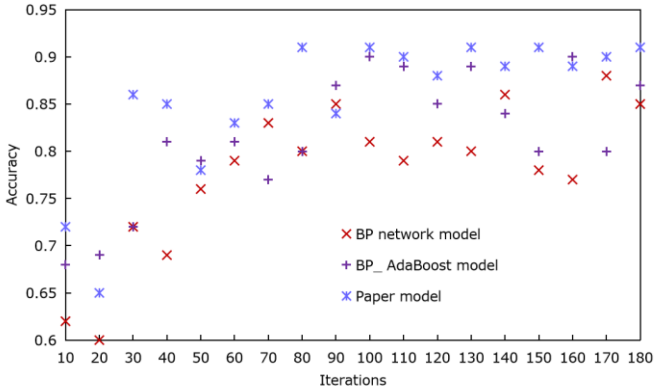

Figure 7: Comparison of accuracy of different identification methods.

Through comparative experiments, it is found that this model has higher accuracy than BP neural network and BP_Adaboost. When the SNR is 0dB, the accuracy of this model reaches 92.97%, which is 5.76% higher than that of BP_Adaboost and 4.98% higher than that of BP neural network. In addition, the model has better stability and robustness, especially in low SNR environment. In this study, a multi-classifier fusion method is used to synthesize the results of multiple base classifiers. This technology can effectively improve the generalization performance of the model and make the model better adapt to various speech signals, thus improving the accuracy and robustness of speech synthesis.

The research's help for AI anchor's computer-aided speech synthesis is mainly reflected in feature extraction, classifier design and multi-classifier fusion. These technologies can effectively improve the accuracy and robustness of speech synthesis and make the generated speech more natural, clear and smooth. In the future, we can further explore how to apply these technologies to the speech synthesis of AI anchors in order to better realize intelligent broadcasting.

6 ACKNOWLEDGEMENT

This work was supported by Hunan Provincial Natural Science Foundation of China with the project "Research on Developing a Human-Machine Symbiotic Hosting Communication Model", project number 2021JJ60021.

Lifen Jiang, https://orcid.org/0009-0001-3444-5126
Xiang Li, https://orcid.org/0009-0000-7073-3646

FERENCES

[1] Abdusalomov, A.-B.; Safarov, F.; Rakhimov, M.; Turaev, B.; Whangbo, T.-K.: Improved feature parameter extraction from speech signals using machine learning algorithm, Sensors, 22(21), 2022, 8122. https://doi.org/10.3390/s22218122

[2] Abreu, A.-F.; Riou, M.; Torrejon, J.; Tsunegi, S.; Querlioz, D.; Yakushiji, K.; Grollier, J.: Role of non-linear data processing on speech recognition task in the framework of reservoir computing, Scientific Reports, 10(1), 2020, 328. https://doi.org/10.1038/s41598-019-56991-x

[3] Ali, W.; Ahmed, A.-A.: Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting, IET Information Security, 13(6), 2019, 659-669. https://doi.org/10.1049/iet-ifs.2019.0006

[4] Alshalan, R.; Al-Khalifa, H.: A deep learning approach for automatic hate speech detection in the saudi twittersphere, Applied Sciences, 10(23), 2020, 8614. https://doi.org/10.3390/app10238614

[5] Chen, L.; Li, S.; Bai, Q.; Yang, J.; Jiang, S.; Miao, Y.: Review of image classification algorithms based on convolutional neural networks, Remote Sensing, 13(22), 2021, 4712. https://doi.org/10.3390/rs13224712

[6] Erdolu, E.: Lines, triangles, and nets: A framework for designing input technologies and interaction techniques for computer-aided design, International Journal of Architectural Computing, 17(4), 2019, 357-381. https://doi.org/10.1177/1478077119887360

[7] Zhang, G.: Quality evaluation of English pronunciation based on artificial emotion recognition and Gaussian mixture model, Journal of Intelligent & Fuzzy Systems, 40(4), 2021, 7085-7095. https://doi.org/10.3233/JIFS-189538

[8] Gao, Z.: Intelligent building BIM fusion data analysis framework based on speech recognition and sustainable computing, International Journal of Networking and Virtual Organisations, 25(1), 2021, 83-101. https://doi.org/10.1504/IJNVO.2021.117760

[9] Jung, Y.-H.; Hong, S.-K.; Wang, H.-S.; Han, J.-H.; Pham, T.-X.; Park, H.; Lee, K.-J.: Flexible piezoelectric acoustic sensors and machine learning for speech processing, Advanced Materials, 32(35), 2020, 1904020. https://doi.org/10.1002/adma.201904020

[10] Kum, S.; Nam, J.: Joint detection and classification of singing voice melody using convolutional recurrent neural networks, Applied Sciences, 9(7), 2019, 1324. https://doi.org/10.3390/app9071324

[11] Li, F.; He, F.; Wang, F.; Zhang, D.; Xia, Y.; Li, X.: A novel simplified convolutional neural network classification algorithm of motor imagery EEG signals based on deep learning, Applied Sciences, 10(5), 2020, 1605. https://doi.org/10.3390/app10051605

[12] Liu, Y.; Li, T.; Zhang, P.; Yan, Y.: LWMD: A comprehensive compression platform for end-to-end automatic speech recognition models, Applied Sciences, 13(3), 2023, 1587. https://doi.org/10.3390/app13031587

[13] Ma, Z.; Liu, Y.; Liu, X.; Ma, J.; Li, F.: Privacy-preserving outsourced speech recognition for smart IoT devices, IEEE Internet of Things Journal, 6(5), 2019, 8406-8420. https://doi.org/10.1109/JIOT.2019.2917933

[14] Pervaiz, A.; Hussain, F.; Israr, H.; Tahir, M.-A.; Raja, F.-R.; Baloch, N.-K.; Zikria, Y.-B.: Incorporating noise robustness in speech command recognition by noise augmentation of training data, Sensors, 20(8), 2020, 2326. https://doi.org/10.3390/s20082326

[15] Savchenko, A.-V.; Savchenko, L.-V.; Makarov, I.: Classifying emotions and engagement in online learning based on a single facial expression recognition neural network, IEEE Transactions on Affective Computing, 13(4), 2022, 2132-2143. https://doi.org/10.1109/TAFFC.2022.3188390

[16] Tu, Y.-H.; Du, J.; Lee, C.-H.: Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(12), 2019, 2080-2091. https://doi.org/10.1109/TASLP.2019.2940662

[17] Vryzas, N.; Vrysis, L.; Matsiola, M.; Kotsakis, R.; Dimoulas, C.; Kalliris, G.: Continuous speech emotion recognition with convolutional neural networks, Journal of the Audio Engineering Society, 68(1/2), 2020, 14-24. https://doi.org/10.17743/jaes.2019.0043

[18] Wang, D.; Wang, X.; Lv, S.: An overview of end-to-end automatic speech recognition, Symmetry, 11(8), 2019, 1018. https://doi.org/10.3390/sym11081018

[19] Xia, K.; Xie, X.; Fan, H.; Liu, H.: An intelligent hybrid–integrated system using speech recognition and a 3D display for early childhood education, Electronics, 10(15), 2021, 1862. https://doi.org/10.3390/electronics10151862