The Application and Prospect of AI Technology in Audio Synthesis

Boan Huang

Tianjin the Second Nankai School, Tianjin, China
Email: andy061012@163.com (B.H.)
Manuscript received September 21, 2024; revised October 27, 2024; accepted November 15, 2024; published December 27, 2024.

***Abstract*—Driven by Artificial Intelligence (AI), speech information processing technology is developing rapidly, among which speech synthesis technology can realize high-fidelity speech output of specified objects and content, and has a wide range of application prospects in human-computer interaction, pan-entertainment and other fields. Traditional audio synthesis methods rely on a lot of manpower and computing power for post-editing and trimming, and the quality of synthesis is not high, which becomes a challenge restricting the development of speech synthesis technology. In this paper, firstly, the sound signal and sound processing are the theoretical basis for audio synthesis. Secondly, the format, parameters and structure of audio files are discussed. Then, the paper reviews the research process and practice methods of neural network in the field of sound synthesis, and takes music as the main research direction. Finally, the application prospect of AI in this field is prospected. Research on the application of AI in the field of sound synthesis will change the traditional method of sound synthesis, improve audio quality and audibility, and promote the innovation and development of audio technology.**

***Keywords*—speech synthesis, audio processing, artificial intelligence, neural networks, application prospects**

I. Introduction

The study of speech synthesis is derived from the human pursuit and demand for language communication. Voice is one of the most natural, convenient, simple and efficient ways of communication in human beings. Through pronunciation, people can express their thoughts, feelings and messages, and interact and communicate with others. Speech synthesis technology is gradually moving into life, and it is of great research significance. It can convert text information into audible sounds, allowing the text to be heard and understood. HMM-Based Speech synthesis technology [1] not only facilitates specific groups in life, but also plays an important role in education, medical care, public transportation and various application fields. It has been widely used in areas such as voice assistants, virtual characters, games and music. The purpose of this paper is to explore the application and prospect of AI technology in sound synthesis, and to analyze its impact on society and industry.

Speech synthesis converts text or text information into audible sounds. It simulates human speech ability and generates natural and smooth synthetic speech through algorithms and models. Sound synthesis can be divided into rule-based methods and statistics-based methods [2]. The rule-based approach uses a speech synthesis engine to simulate the movement of human articulatory organs to produce realistic speech sounds. The statistics-based method uses machine learning algorithms to learn the characteristics and patterns of sound from a large number of speech data, and then generate speech. The continuous progress of these technologies makes the sound synthesis closer and closer to the real human sound, providing a better experience for human-computer interaction.

Artificial intelligence has had a profound impact on speech synthesis technology, bringing its development into a new stage. First, deep learning methods are not limited by the language. Through end-to-end training, deep learning technology can easily adapt to the speech synthesis needs of different languages through language differences. Secondly, deep learning methods have achieved great improvement in the quality of synthetic speech. Traditional statistical methods usually require complex feature engineering and rule design, while deep learning can automatically learn the mapping relationship of speech features through a large number of data and neural network models. This generates a more natural, smooth, synthetic speech with accurate intonation and emotional expression. In addition, deep learning also brings about a simplification in the process of speech synthesis. Deep learning method can directly from input text to speech generation, the whole process is more concise and efficient, which greatly reduces the generation time of synthetic speech. Due to these significant advantages of deep learning method, it has become the mainstream of speech synthesis technology and become a research hotspot for scholars.

By adopting the method of artificial intelligence, the speech synthesis technology has achieved a significant improvement in terms of rhythm and naturalness, and has greatly increased its application value. An important application scenario is the film and television dubbing, especially the cross-language dubbing. This not only saves a lot of time and cost, but also can maintain a better synthesis effect, so that the film and television works can be spread globally faster. Another key area of application is music synthesis. By combining artificial intelligence and speech synthesis technology, songs of virtual singers can be generated, including elements such as lyrics, melodies and sounds. This brings more possibilities for music creation, allowing music producers to get rich creative inspiration in the generation process and use it in advertising, games, entertainment and other fields. In addition, audiobooks use speech synthesis technology to transform text content into audio books with phonetic expression. This allows people with hearing disorders or visual impairment to access the content of books, providing easier access to knowledge and culture.

In conclusion, the application and prospect of AI technology in sound synthesis is a research field of much attention. Through continuous innovation and development, sound synthesis will be more widely used in various fields, bringing better experience and services to people. However,

sound synthesis techniques also face some challenges that need to be further studied and addressed.

II. The Basic and the Process of Sound Processing Technology

There are several methods and models for speech synthesis technology. One commonly used method is rule-based speech synthesis [3], which utilizes predefined acoustic and linguistic rules to synthesize speech. Although this method offers high control and interpretability, it is limited by the need for extensive manual design and annotation. Waveform concatenation is a commonly used technique in speech synthesis [4]. It involves concatenating units of data, such as phonemes, from pre-recorded human speech samples to create a complete synthesized speech waveform. The selection of the units to be concatenated can be dynamically adjusted based on the input text content and context, resulting in more accurate and natural-sounding synthesized speech.

To overcome the limitations of rule-based speech synthesis, statistical modeling and machine learning methods have been extensively employed. Among them, statistical parametric speech synthesis utilizes abundant speech training data and related features to predict acoustic parameters, generating natural and fluent synthesized speech. This method has achieved significant success, with the introduction of deep neural networks further enhancing the quality of synthesized speech. According to Ling at al. [5], Hidden Markov Models (HMMs) are statistical models widely used in speech synthesis. In speech synthesis, HMMs are employed to model the speech signal by learning the probability distribution of speech features, enabling the prediction of acoustic parameters for synthesized speech. One major advantage of HMMs is their ability to capture the temporal and dynamic characteristics of speech, resulting in more coherent and natural-sounding synthesized speech.

Additionally, a more recent approach is deep learning-based speech synthesis [6]. With deep neural network models, this method directly learns the mapping relationship between input text and speech features, producing natural and fluent synthesized speech. It simplifies the training process in traditional speech synthesis and accomplishes better synthesis results. Deep Neural Network (DNN)-based speech synthesis utilizes deep neural networks to model the process of speech generation. By training multi-layer neural networks, this method automatically learns the mapping between input text and speech features, generating synthesized speech that sounds more natural and fluent. The introduction of DNNs has significantly improved the accuracy and realism of speech synthesis. Recurrent Neural Network (RNN)-based Speech Synthesis: RNNs, a type of neural network with recurrent connections, are widely used in speech synthesis. RNNs excel at processing sequence data and maintaining memory of previous information to model the current input. By establishing an RNN model, speech synthesis systems can better capture the contextual information and temporal dependencies of speech, enhancing the coherence and fluency of synthesized speech [7]. Long Short-Term Memory (LSTM)-based speech sythesis: LSTM is a variant of RNNs that addresses the issue of vanishing or exploding gradients when dealing with long sequence data. Through the introduction of gating mechanisms, LSTM has been widely applied in speech synthesis to model long-term dependencies and handle long-text inputs, resulting in improved accuracy and naturalness of synthesized speech [8].

A. Sound Signal

To begin with, the basic of sound processing technology is the sound signal. It refers to the sound produced by musical instruments or recorded audio [9]. It exhibits periodic characteristics that can be perceived by our auditory system. From the physics aspect, it includes timbre, pitch, loudness. Timbre refers to the specific quality or perception of sound. Different instruments and voices have unique timbres. Timbre is determined by the harmonic and resonant characteristics in the audio signal [10]. Pitch refers to the frequency of a sound, which determines whether a sound is perceived as high or low. Pitch is directly related to frequency, where higher frequencies correspond to higher pitches. Loudness refers to the intensity or volume of a sound, determining whether a sound is perceived as loud or soft. Loudness is related to the amplitude of the sound signal, where greater amplitudes result in higher loudness [11]. These aspects form the sense of auditory.

B. Sound Processing

Sound processing is the process of applying various operations to an audio signal [12]. It includes gain control, filtering, sampling, conversion, processing, and synthesis. These operations are used to modify the volume, timbre, spatial characteristics of the sound, remove noise, apply special effects, and generate new sounds.

Gain Control: Gain control involves adjusting the volume or amplifying the amplitude of a sound signal shown in Fig. 1. It can be used to increase the loudness of an audio signal or reduce noise in the signal.

Fig. 1. Gain control.

Filtering: Filtering involves altering the frequency characteristics of an audio signal to adjust its timbre or remove noise. Common filters include low-pass filters, high-pass filters, and band-pass filters shown in Fig. 2.

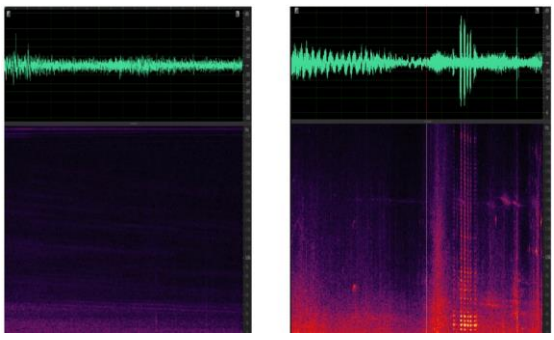

Fig. 2. The common filters.

Sampling: Sampling refers to converting a continuous analog sound signal into a discrete digital signal, as shown in Fig. 3. By taking periodic samples of the continuous signal, sound signals can be converted into a digital form for processing and storage.

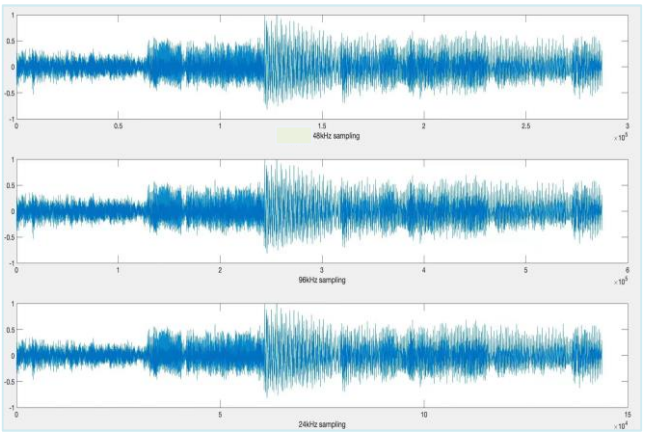

Fig. 3. The example of sampling.

Transcoding: Conversion involves transforming a sound signal from one format or encoding to another. For example, converting analog sound signals to digital audio files (e.g., MP3) or converting digital audio signals back to analog sound signals.

Processing and Analyzing: Sound processing involves applying various algorithms and effects to audio signals to modify their timbre, spatial characteristics, or create special audio effects.

Synthesis: Sound synthesis involves generating new sounds using synthesizers, synthesis software, or other tools. Synthesis techniques can emulate instrument sounds, produce vocal sounds, or synthesize electronic sound effects. (Fig. 4)

Fig. 4. The synthesizer was used to harmonize.

Sound processing is related to audio files. An audio file is a digital representation used to store sound signals. Sound processing involves various operations and manipulations of audio files to achieve specific effects or alter their characteristics. Through sound processing, audio files can undergo operations such as gain control, filtering, equalization adjustments, time-domain processing, and frequency-domain processing. These processes can change the volume, timbre, spatial perception, and dynamic range of the audio, remove noise, or add special effects. Sound processing involves adjusting the sample rate and bit depth of audio files, converting formats, and synthesizing new sound elements. Therefore, sound processing is the process of modifying and optimizing audio files to achieve desired effects.

C. Audio File Format

Audio files have important definitions and functions in sound signal processing. They serve as the input and output carriers for sound signal processing algorithms, allowing for analysis, processing, and reproduction of the sound signals through their digital representations. Specifically, audio files have the following roles in sound signal processing: Data recording and storage, Signal capture and acquisition, Signal transmission and sharing, Signal processing and analysis. The audio files contain the format, parameters and structures.

1) Audio file

An audio file format refers to a specific file format used for storing audio data . It determines how the data is organized, encoded, and compressed within the audio file. Common audio file formats include WAV, MP3, AAC, FLAC.

- **WAV** (Waveform Audio File Format) is a lossless audio file format that stores audio data using PCM (Pulse Code Modulation) encoding. WAV files are typically large in size and high in quality as they are uncompressed. It also supports various sample rates and bit depths, making it suitable for professional audio recording and editing.
- **MP3** (MPEG-1 Audio Layer 3) is a lossy audio file format that achieves high compression ratios by removing less perceptible audio details, resulting in smaller file sizes. MP3 format is widely used in music players and for internet streaming due to its smaller file size and good audio quality.
- **AAC** (Advanced Audio Coding) is a lossy audio file format that uses advanced compression algorithms to achieve higher audio quality and smaller file sizes compared to MP3. AAC format is widely used in music downloads, streaming media, and mobile devices.
- **FLAC** (Free Lossless Audio Codec) is a lossless audio file format that reduces the file size of audio files using lossless compression algorithms while preserving audio quality. FLAC format is suitable for scenarios where high-quality audio needs to be retained, such as music production and audio archiving.

Table 1. Common audio file format

Audio format	Compression algorithm	Sound quality	File size	Common extension
MP3	Lossy compression	Medium	Small	.mp3
WAV	Lossless compression	High	Large	.wav
FLAC	Lossless compression	High	Medium	.flac
AAC	Lossy compression	Medium	Small	.aac
OGG	Lossless compression	Low	Low	.ogg

In addition to the commonly used audio file formats mentioned above, there are many other formats such as OGG, WMA, ALAC, each with its own characteristics and suitable applications, showing in Table 1. Choosing the appropriate audio file format depends on application requirements, including audio quality, file size, device compatibility, and network transmission considerations.

2) Audio file parameters

Audio file parameters are a set of parameters used to describe the characteristics and properties of audio data [13]. These parameters provide basic information about the audio file, such as sample rate, number of channels, bit rate. Below is a detailed explanation of some common audio file parameters:

Sample Rate: The sample rate refers to the number of times the audio signal is sampled per second [14]. It represents the accuracy and time resolution of capturing the audio signal within a unit of time. Common sample rates include 44.1 kHz, 48 kHz, with 44.1 kHz being the standard sample rate for CD audio quality.

Channels: Channels represent the number of simultaneous recording or playback channels for the audio signal. Commonly used channel configurations include mono (single channel) and stereo (left and right channels). There are also more advanced multi-channel configurations such as 5.1 channel, 7.1 channel.

Bit Rate: The bit rate refers to the number of bits transmitted per second and represents the compression level of the audio data [15]. A higher bit rate generally indicates better audio quality but also results in larger file sizes. Common bit rates include 128 kbps, 320 kbps.

Compression Format: The compression format refers to the method and encoding algorithm used for compressing the audio data. Common compression formats include MP3, AAC, WMA, etc., which reduce file sizes by discarding or simplifying less perceptible audio details.

Sample Size: Sample size refers to the number of bits used to represent each sample in the audio data [16]. Common sample sizes include 8-bit, 16-bit, 24-bit, etc. A higher bit depth increases the dynamic range and accuracy of the audio.

These parameters are primarily used to describe the basic characteristics of audio files, and different file formats and applications may have different requirements. Proper selection and understanding of these parameters ensure audio quality and compatibility in audio recording, storage, and transmission.

3) Audio file structure

The structure of an audio file refers to the organization method and format used for storing audio data [17]. An audio file is composed of multiple parts, including header information, audio data, and footer information. The specific file structure can vary depending on the file format, with common audio file formats including WAV, MP3, AAC. Generally, the structure of an audio file is as follows:

Header Information: This section contains metadata about the file, such as file type, sample rate, number of channels, bit rate. The header information is used by players or decoders to correctly parse the audio data.

Audio Data Chunk: This section actually stores the audio signal data. The audio data chunk contains a continuous stream of audio samples, which may be stored in different encoding formats and compression algorithms depending on the specific file format.

Footer Information: This section contains additional information or end-of-file markers. Footer information is often used to assist players or decoders in correctly identifying the end position of the audio file.

Different audio file formats may have different structures and organization methods. For example, WAV files use lossless compression to store audio data, with the header information containing basic parameters of the audio and the audio data being stored directly in PCM (Pulse Code Modulation) format. On the other hand, audio files in formats like MP3 and AAC use complex algorithms to compress and encode the audio data in a lossy manner, reducing the file size. In Fig. 5, the structure of files in MP3 are shown.

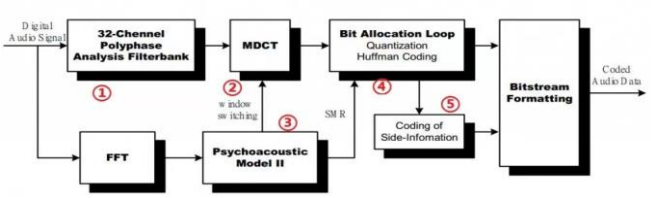

Fig. 5. The process of MP3 files.

III. The Application and Prospect of AI in Audio Synthesis

From the previous section, we can see that these techniques play important roles in the field of speech synthesis and have made significant advancements. With ongoing developments and in-depth research, we can expect more accurate, natural, and human-like synthesized speech generation. In the discussion, we will focus on the application and prospect of AI technology in this part, starting with the composition of music as well as the process of sound.

A. Application of AI Technology in Sound Technology - Take Music as An Example

AI technology can be used for music generation, music analysis and music recommendation. In terms of music generation, AI can generate new music works by learning a large number of music works and patterns, and combining algorithms and models. This technology helps music creators get inspired and create new pieces.

1) Data collection and preprocessing

Data collection and preprocessing are important steps in the application of AI technology in sound technology. Here are the general steps for data collection and preprocessing in the context of AI technology applied to music:

Data Source Selection: Determine the source of the data, which can be publicly available datasets, data from professional music platforms, or data provided by artists. Ensure that the chosen data sources are reliable and legally owned.

Data Type Selection: Determine the type of data needed, such as audio files, song metadata, sheet music, lyrics, etc. Different applications may require different types of music data.

Data Filtering and Cleaning: Perform initial screening of collected data to exclude incomplete or poor-quality data. Additionally, clean the data to remove noise, interference, or

other unwanted elements.

Data Annotation and Labeling: Annotate and label the collected music data, adding information such as track details, note/chord annotations, emotion tags, etc., for subsequent analysis and processing.

Feature Extraction: Extract meaningful features from the audio data, such as pitch, timbre, rhythm, volume, etc. These features will serve as inputs to subsequent machine learning algorithms.

Data Partitioning and Normalization: Divide the collected data into training, validation, and testing sets for model training, validation, and evaluation. Normalize the data to ensure consistency and comparability.

Data Format Conversion: Convert the music data into appropriate formats based on the requirements of the used machine learning algorithms and tools, such as digital signals, spectrograms, or other specific input forms.

Data Augmentation: To increase the diversity and generalization ability of the data, data augmentation techniques can be applied during the preprocessing stage, such as randomly altering the pitch, speed, or adding noise to the audio.

Data Storage and Management: Establish a suitable database or data storage system that enables quick access and management of the music data, while ensuring data security and confidentiality.

These steps provide a general framework, and the specific methods for data collection and preprocessing may vary depending on the specific requirements and available resources of the application.

2) Feature engineering

In the field of feature engineering, RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), and GAN (Generative Adversarial Network) are three commonly used deep learning models. These models have unique application scenarios and advantages in feature engineering and are widely used in areas such as speech recognition, image processing, and natural language processing [18].

RNN (Recurrent Neural Network)

RNN is a sequence model that incorporates recurrent connections in neural network structures to process sequential data. Unlike traditional feedforward neural networks that cannot retain past inputs, RNN's recurrent connections allow the current output to be correlated with previous states, thus capturing temporal information in time series data. RNN can be used for prediction, classification, and generation of variable-length sequences [19].

The basic neural network only establishes weight connections between layers, while the biggest difference of RNN is that the weight connections are also established between neurons between layers, as shown in the Fig. 6. This is a standard RNN structure diagram, where each arrow represents a transformation, i.e., the arrow connection has a weight. The left side is folded up, the right side is unfolded, and the arrow next to the 'h' on the left represents the 'loop' in this structure embodied in the hidden layer.

In the expanded structure, we can observe that in the standard RNN structure, there are also weights between the neurons in the hidden layer. That is, as the sequence progresses, the hidden layers in the front will affect the hidden layers in the back. In the figure, 'O' represents the output, 'y' represents the determined value given by the sample, and 'L' represents the loss function. We can see that the 'loss' also accumulates with the recommendation of the sequence.

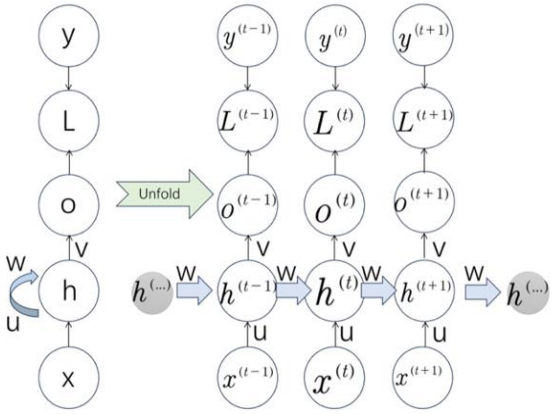

Fig. 6. Standard RNN structure diagram.

CNN (Convolutional Neural Network)

CNN is a neural network model applicable to image and spatial data. It utilizes convolution and pooling operations to extract local and global spatial features, enabling deep hierarchical feature representation learning through multiple convolutional and pooling layers. CNN effectively captures spatial relations and features in images, exhibiting excellent performance in tasks such as image classification, object detection, and image generation.

The training process of convolutional neural networks (CNN) is composed of two stages - the forward propagation stage and the backward propagation stage. During the forward propagation stage, the input data flows from the low-level layers to the high-level layers in the network. Then, if the output of the network differs from the expected result, the error is backpropagated from the high-level layers to the low-level layers during the backward propagation stage [20]. This process continues until the error is minimized or reaches an acceptable level. The training process proceeds as follows:

- The network initializes the weights.
- The input data is fed forward through the convolutional layer, subsampling layer, and fully connected layer to obtain the output value.
- The difference between the network's output value and the target value is calculated.
- If the error is greater than our desired tolerance level, the error is backpropagated to the network. In turn, the error of each layer is obtained: the fully connected layer, the subsampling layer, and the convolutional layer. The error of each layer can be understood as how much the network should bear for the total error of the network.
- The weights are updated based on the calculated error. Then we proceed to step two.

The total flow of the process is shown in Fig. 7.

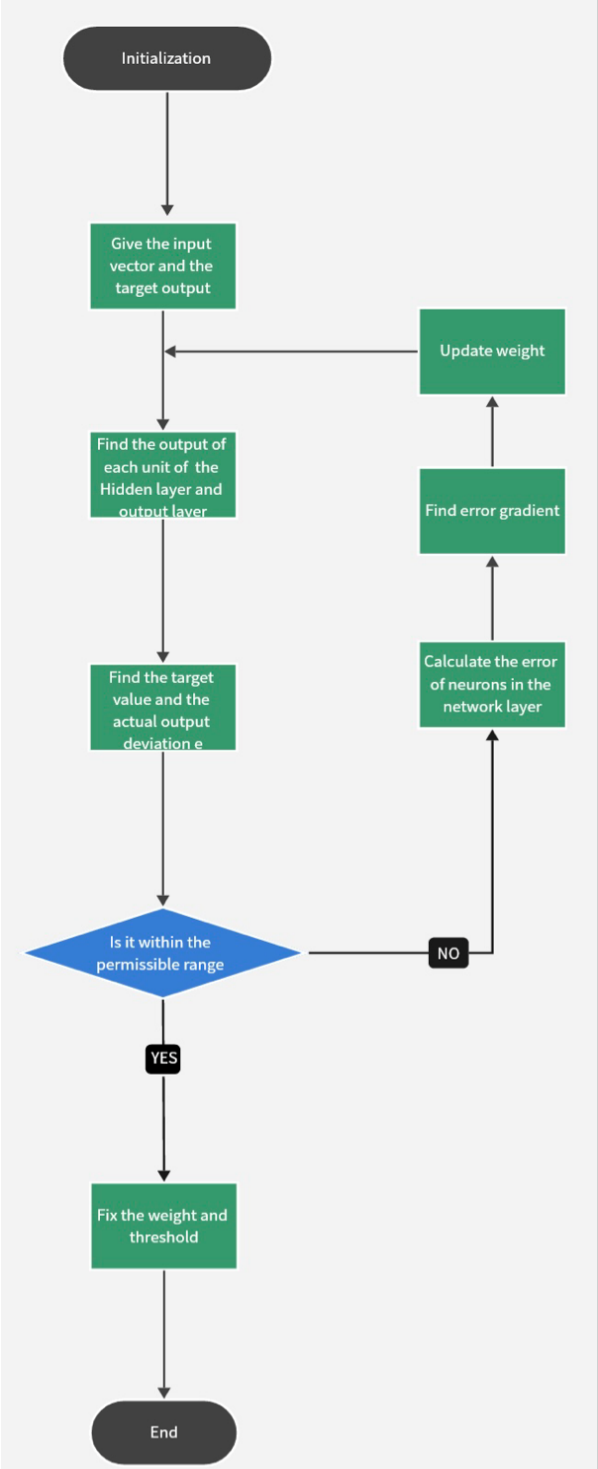

Fig. 7. The flow chart of CNN.

GAN (Generative Adversarial Network)

GAN is an adversarial model composed of a generator and a discriminator. The generator attempts to generate samples that appear realistic, while the discriminator aims to differentiate between samples generated by the generator and real samples. Through a dynamic adversarial training process, the generator continuously improves the quality of generated samples, while the discriminator enhances its ability to discern between them. GAN has powerful capabilities in generating realistic new samples, such as images, audio, or text, and excels in feature expression and pattern learning for generating data [21]. The whole process of GNN is shown in Fig. 8.

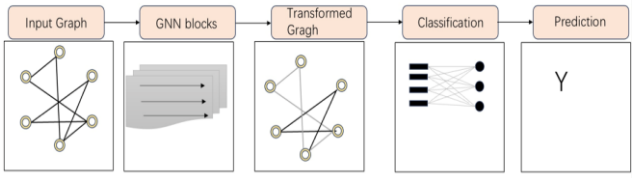

Fig. 8. The flow chart of GNN.

The specific network structure and update iteration formulas in each layer are considered in detail below:

$$h_v^{(k+1)} = \sigma\left(W_k \sum_{u \in N(v)} \frac{h_u^{(k)}}{|N(v)|} + B_k h_v^{(k)}\right), \forall k \in \{0, \ldots, \ k-1\}$$

Initial layer for node input characteristics is: $h_v^0 = x_v$, and the last layer is the result of embedding: $z_v = h_v^{(k)}$. In the iterative formula, σ is a nonlinear activation function (e.g., ReLU). The middle two parts of the activation function are: the average embedding of the neighborhood nodes in the previous layer is multiplied by a trainable weight W_k ; And the embedding of the current node in the last iteration $h_v^{(k)}$ multiplied by the trainable weight B_k.

$$\text{Let } H^{(k)} = \left[h_1^{(k)} \ldots h_{|v|}^{(k)}\right]^t, \text{then} \sum_{u \in N_v} h_u^{(k)} = A_v, : H^{(K)}$$

where A_V is the v row of the adjacency matrix A. Then construct an angular matrix D, where, $D_{v,}v = Deg(v) = |N(v)|$, the reverse D^{-1} can be expressed as: $D_{v,}v = Deg(v) = |N(v)|$. Then the transfer term of the neighbor node in the previous iteration formula can be expressed as a matrix form:

$$\sum_{u \in N(v)} \frac{h_u^{(k-1)}}{|N(v)|} \rightarrow H^{(K+1)} = D^{-1} A H^{(k)}$$

The final iterative formula in matrix form is:

$$H^{(K+1)} = \sigma\left(A H^{(k)} W_K^T + H^{(k)} B_k^t\right)$$

(A) (B)

where (A) and (B) correspond to the information transmission of neighbor nodes and their own information transmission respectively, as shown in the Fig. 9 below:

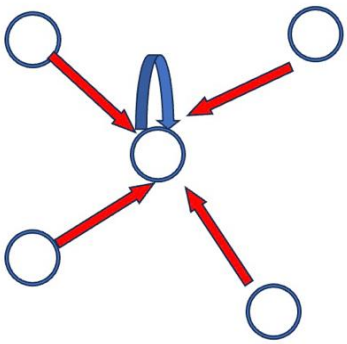

Fig. 9. The information transmission of neighbor nodes and their own information transmission.

3) Training

Data Collection and Sampling: Collect and obtain data in related fields according to the problems that need to be solved. This data can be labeled or unlabeled. Samples are then taken according to the characteristics of the data to ensure that the training set is representative of the distribution of the real data.

Model Selection: Select the appropriate neural network model according to the nature of the problem and the characteristics of the data. Common models include deep feedforward neural networks, convolutional neural networks, recurrent neural networks, etc. According to the complexity of the problem and the size of the data, choose the appropriate model structure and number of layers.

Model Splicing: If the problem to be solved is more complex, it may be necessary to combine multiple models to build a composite model. This can be done by stacking models, series models, or parallel models. Model splicing can improve the expressiveness and prediction performance of models.

Design Model: According to the selected model structure, design each layer of the network, activation function, loss function, etc. When designing the model, you need to consider the shape and type of input data, and set the appropriate hyperparameters according to the requirements of the problem, such as learning rate, batch size, etc. In addition, you also need to consider whether to use regularization, optimization algorithms, initialization strategies, etc. For the part of loss function, there are two kinds of losses, supervised losses (using the label of the node) and unsupervised losses(using the structure of the graph), where the supervised loss is the common one:

$$\min_{\theta}(y, f(z_v))$$

y is the label of the node, and f (zv) indicates another function, such as sigmoid, soft max function, etc. C is the L2 loss, cross entropy loss, etc. [22]. For the part of activation function, soft max function is widely used [23]:

$$S_i = \frac{e^i}{\sum_j e^j}$$

The graph of this function is shown in Fig. 10. The process could be divided into three layers, which are the input layer, neuron layer and the output layer, shown in Fig. 11.

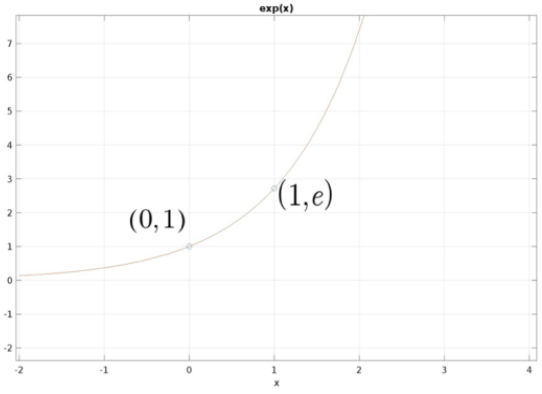

Fig. 10. The graph of the function.

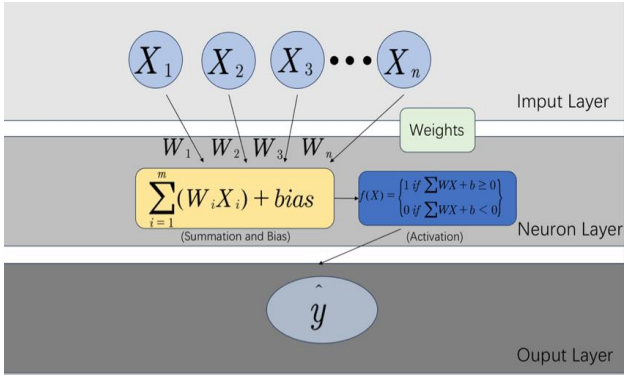

Fig. 11. The process could be divided into three layers.

B. Assessment

The following methods can be used to assess neural network models:

Loss Function: An appropriate loss function is used to measure the error of the model on the training data. Common loss functions include Mean Square Error and Cross-Entropy. A lower loss value indicates that the model performs better on the training data.

Accuracy: Calculate the classification accuracy of the model on the verification or test set. Accuracy is the ratio of the number of samples that the model predicts correctly to the total number of samples. Higher accuracy means that the model performs better in sample classification.

Precision and Recall: For multi-class classification problems, precision and recall can be used to evaluate the performance of the model. The accuracy rate is the ratio of the number of samples correctly predicted for a category to the number of all samples predicted for that category. The recall rate is the ratio of the number of samples correctly predicted for a category to the number of all samples in that category.

F1-Score: For the problem of unbalanced category distribution, there may be a compromise between accuracy and recall. F1-Score combines accuracy and recall to measure the performance of the model. F1-Score is the harmonic average of accuracy and recall.

Cross-Validation: By dividing the data set into multiple subsets for training and validation, the stability and generalization ability of the model can be evaluated. Common cross-validation methods include K-fold cross-validation and leave-one cross-validation.

All of the above indexes can be used to evaluate the performance of the neural network model, select the evaluation indexes suitable for the task requirements, and conduct comprehensive evaluation combined with the actual situation to find the optimal model.

IV. Conclusion

The main goal of the current study was to determine the application and potential of AI technology in audio synthesis. This study has identified that the application of artificial intelligence technology in the audio field has made remarkable progress. The processing and analysis of sound signal is an important prerequisite for intelligent speech interaction and audio recognition. Furthermore, sound processing technology can be used to provide in-depth

analysis and understanding of sound signals to support a variety of applications. Besides, various neural network models such as RNN, CNN and GAN are widely used in music recommendation, music generation and other tasks. In the field of music, the application of artificial intelligence technology can not only improve the accuracy and efficiency of music recommendation and music generation, but also help us better understand and analyze music works. At the same time, the application of this technology also provides us with more ways to experience music and richer music content.

The insights gained from this study may be of assistance to the prospect of AI audio synthesis. The application of artificial intelligence technology in the field of sound technology is changing the way we understand and apply sound signals. In the future, with the continuous progress of technology and the continuous expansion of application scenarios, we have reason to believe that the application of this technology will be more extensive and in-depth.

CONFLICT OF INTEREST

The author has claimed that no conflict of interest exists.

REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *Transactions of the Institute of Electronics Information & Communication Engineers*, vol. 83, no. 3, pp. 2099–2107, 1999.

[2] Z. Wei, D. X. Yu, and Y. C. Yan, "A method combining rule-based and statistics-based approaches for Chinese word segmentation," *Application Research of Computers*, vol. 21, pp. 3, pp. 23–25, 2004.

[3] M. Eric and C. Francis, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5–6, pp. 453–467, 1989.

[4] Y. Arai, R. Mochizuki, H. Nishimura, and T. Honda, "An excitation synchronous pitch waveform extraction method and it's application to the VCV-concatenation synthesis of Japanese spoken words," *Proc Icslp*, vol. 3, pp. 1437–1440, 1996.

[5] Z. Ling, Y. Wu, Y. Wang, L. Qin, and R. Wang, "USTC system for blizzard challenge 2006 an improved HMM-based speech synthesis method," *Computer Science*, vol. 4, 2006.

[6] O. Watts, E. G. Henter, T. Merritt, Z. Wu, and S. King, "Listening test materials for 'From HMMs to DNNs: Where do the improvements come from?'" *Technology Research (CSTR), IEEE*, pp. 5505–5509, 2016.

[7] A. R. Mohamed, A. Graves, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

[8] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 4470–4474.

[9] M. Shashoua and D. Glotter, "Method and system for enhancing quality of sound signal," *Acoustical Society of America (ASA)*, vol. 107, no. 6, pp. 2946, 1999.

[10] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2595–2603, 1993.

[11] P. Gramming, J. Sundberg, M, S. Ternstr, R. Leanderson, and W. H. Perkins, "Relationship between changes in voice pitch and loudness," *Journal of Voice*, vol. 2, no. 2, pp. 118–126, 1988.

[12] P. Kiljan, W. Moczulski, and K. Kalinowski, "Initial study into the possible use of digital sound processing for the development of automatic longwall shearer operation," *Energies*, vol. 14, no. 10, pp. 2877, 2021.

[13] G. Aditi, R. Ranjan, M. Rahul, and A. Shivani, "Methods and systems for adjusting audio parameter," India, Patent 385223, December 19, 2021.

[14] C. Schmidmer, R. Bitto, and M. Keyhl, "Device and method for determining a sample rate difference," U.S. Patent 9,037,435, issued May 19, 2015.

[15] J. Beran and R. Sherman, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans Commun*, vol. 43, no. 2, pp. 1566–1579, 1995.

[16] R. C. MacCallum, M. W. Browne, and H. M. Sugawara, "Power analysis and determination of sample size for covariance structure modeling," *Psychological Methods*, vol. 1, no. 2, pp. 130, 1996.

[17] M. T. Hakkinen and G. Kerscher, "Structured audio: Using document structure to navigate audio information," *Presentation at the CSUN Conference on Technologies and Disabilities*, 1998.

[18] A. Severyn and A. Moschitti, "Automatic feature engineering for answer selection and extraction," in *Proc. the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 458–467, 2013.

[19] T. Lei and Y. Zhang, "Training RNNS as fast as CNNS," 2017. Available: https://openreview.net/forum?id=rJBiunlAW

[20] S. Ma, S. A. Shugao *et al.*, "Do less and achieve more: Training CNNS for action recognition utilizing action images from the web," *Pattern Recognition*, vol. 68, pp. 334–345, 2017.

[21] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "Gan augmentation: Augmenting training data using generative adversarial networks," *arxiv preprint arxiv*:1810.10863, 2017.

[22] W. Liu, Y. Wen, Z. Yu and M. Yang, "Large-margin softmax loss for convolutional neural networks," *arxiv preprint arxiv*:1612.02295, 2016.

[23] A. De Brebisson and P. Vincent, "An exploration of softmax alternatives belonging to the spherical loss family," *arxiv preprint arxiv:*1511.05042, 2015.