Speaker Age Recognition based on Convolution and Transformer Fusion Framework

Zheyan Zhang[0000-0001-5753-4853], Renwei Li[0009-0004-5828-5069], Kewei Chen*[0000-0002-3106-4336]

Faculty of Mechanical Engineering & Mechanics, Ningbo University, Ningbo 315211, China
{2201090021, 2201090013, dongfangyan, chenkewei} @nbu.edu.cn
*Corresponding author: chenkewei @nbu.edu.cn

Abstract. In the process of human-computer intelligent speech interaction, inferring the speaker's age from human speech signals presents a challenging task. The acoustic features related to age in speaker's voice are complex, making it difficult for traditional machine learning methods to achieve comprehensive and accurate recognition results. This paper proposes a method for speaker age recognition based on a framework that integrates convolutional and self-attention mechanisms. Firstly, speech signals are transformed into spectrograms, and a CNN-Transformer Dual Branch Parallel Fusion Network (CTPF-Net) is designed to achieve a comprehensive extraction of global and local detail features of speech signals. Additionally, gender information is considered during training to perform unified age-gender recognition, achieving better accuracy than age recognition alone. Experiments and analysis on the Common Voice dataset demonstrate that the proposed model achieves an average accuracy of 84.5% in age recognition tasks. Moreover, without significantly increasing model complexity, the model can accurately differentiate speakers across different age segments.

Keywords: Speaker Age Recognition; Spectrogram; Convolutional Neural Network; Self-Attention Mechanism; Feature Fusion.

1 Introduction

Speech is considered the most convenient mode of human-computer interaction for information delivery [1]. With the rapid development of artificial intelligence technology, expectations for human-computer intelligent speech interactions have increased. Current research not only focuses on the semantic information provided by speech but also shifts attention to other embedded information such as the speaker's identity, emotions, gender, and even age [2]. Endowing machines with the ability to recognize speaker age and providing age-appropriate personalized content and services can greatly enhance the emotional experience and application efficiency of users.

Speaker age recognition typically involves the analysis of complex signals whose features may be variable, nonlinear, and difficult to interpret. This necessitates the

development of sophisticated models and algorithms to identify and extract relevant features from signals characterized by significant individual differences and variability. There are various methods for extracting acoustic features, but the most commonly used in speaker age recognition are prosodic features, spectral, and cepstral-based features. Yue et al. [3] utilized isolated word speech with support vector machines and Mel frequency cepstral coefficients to recognize ages in youth, middle-aged, and elderly groups. Chen et al. [4] considered the impact of the speaker's emotional state on age recognition, combining speech parameters such as fundamental frequency, zero-crossing rate, and Mel frequency cepstral coefficients under different emotional states, and constructed a speaker age recognition system based on the Gaussian Mixture Model (GMM). Du et al. [5] proposed a statistical analysis recognition method based on multi-resolution features of effective frequency bands. This method employs wavelet packet transform to decompose audio into effective frequency bands, connecting wavelet packet coefficients of each band to form a comprehensive calculation of Mel frequency cepstral coefficients, obtaining multi-resolution feature parameters (WPMFC), and modeling using Gaussian Mixture Models. Additionally, Bahari et al. [6] have proposed a new method for estimating speaker age based on i-vectors, where each utterance is modeled by its corresponding i-vector. Intra-speaker covariance normalization techniques are then used to compensate for session variability, and finally, Least Squares Support Vector Regression (LSSVR) is applied to estimate the speaker's age.

Despite some progress in research on speaker age recognition, extracting significant speech age characteristics and designing high-performance classification models remain challenges. The similarity of temporal and spectral acoustic characteristics across different age groups is one of the reasons for the lower accuracy when using acoustic features for age classification [7,8]. Therefore, many scholars have adopted deep learning methods, directly inputting temporal, frequency domain representations, or spectrograms into neural networks, utilizing the powerful data fitting capabilities of deep networks to extract latent features for analysis. Ghahremani et al. [9] applied the x-vector neural network architecture for speaker age recognition, mapping variable-length utterances to fixed-dimension embedding vectors containing relevant sequential information to construct x-vectors. These x-vectors are then used to estimate age based on the speaker's speech signals. Tursunov et al. [10] designed a Multi-Attention Module (MAM) that considers both spatial and temporal features of speech signals, which was combined with Convolutional Neural Networks (CNN) to enhance the ability to classify age and gender. Sánchez-Hevia et al. [11] proposed a secure and automatic speech recognition system named LimitAccess, which distinguishes between children and adult voices using spectrograms and MFCCs through a lite CNN model. Mavaddati [12] tested a series of CNNs and RNNs for performance in a speech age recognition system and proposed a fine-tuned ResNet34 architecture combined with data augmentation and transfer learning to improve robustness and performance on new data.

This paper proposes a method for speaker age recognition based on a framework integrating convolutional and self-attention mechanisms, using a designed CNN and Transformer Dual Branch Parallel Fusion Network (CTPF-Net) to extract high-dimensional global and local features from spectrograms. By incorporating gender information into age recognition, the accuracy of age recognition has been improved.

Experiments on the Common Voice dataset and comparisons with existing methods validate the effectiveness of this model.

2 Proposed Age recognition Methodology

Temporal and frequency domain analyses play crucial roles in speech processing, yet both have limitations when used independently. Temporal analysis lacks an intuitive understanding of the frequency characteristics of speech signals, while frequency domain analysis lacks information on how speech features evolve over time [13]. Since speech is a time-varying signal, its spectrum also changes over time. Therefore, the age recognition method proposed in this paper is based on the analysis of spectrograms. Spectrograms integrate the characteristics of frequency spectra and temporal waveforms, vividly displaying how the speech spectrum changes over time and containing a wealth of information about the speaker, which is commonly used in voiceprint recognition. Additionally, converting speech signals into two-dimensional images allows for the application of advanced algorithms from image processing, enhancing the models for speech analysis and processing.

As illustrated in Figure 1, the proposed speech age recognition process begins by converting speech signals into spectrograms. Subsequently, the designed CTPF-Net model extracts deep features from both local and overall aspects of the spectrogram. Finally, the age of the speaker is determined through a classification model.

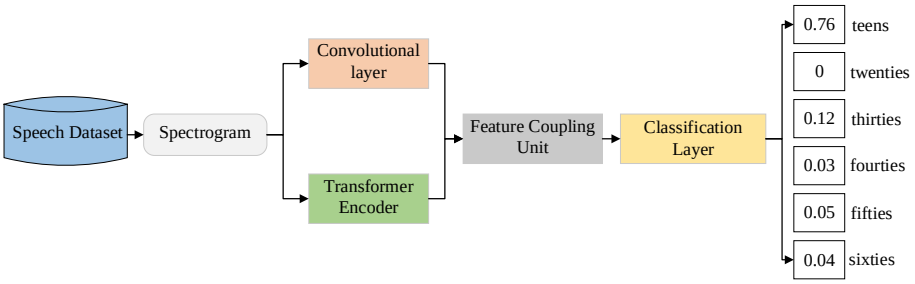

Fig. 1. Speech Age Recognition Process

2.1 CNN and Transformer parallel fusion network (CTPF-Net)

CTPF-Net, as depicted in Figure 2, consists of two parallel processing branches: a CNN branch and a Transformer encoder branch. The CNN branch includes three convolutional layers (Conv), two batch normalization layers (BN), and two pooling layers (MP). The first convolutional layer captures local features in the spectrogram, the second serves as an intermediate layer to generate feature maps for subsequent operations, and the third convolutional layer extracts deeper, hidden cues from the input data. Through this CNN branch, the model learns the implicit relationships between time and frequency in the spectrogram to obtain relatively significant local age characteristics.

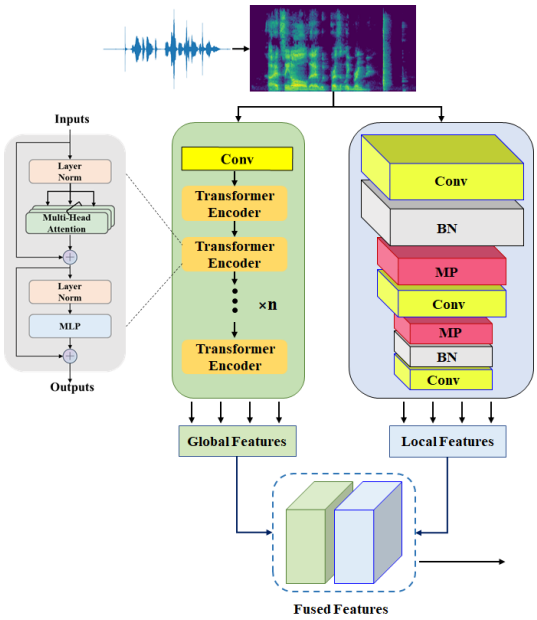

Fig. 2. Structure of CTPF-Net

The Transformer, known for its self-attention mechanism [14], has robust parallel computation capabilities and can learn long-range dependencies between image pixels on a global scale, thereby possessing strong global information extraction abilities. This paper adopts the encoder portion of the Transformer and constructs a Transformer encoder branch. Initially, convolution operations uniformly segment the image into N blocks, followed by a flattening operation to convert it into a two-dimensional matrix format. This matrix is then fed into a D_0 dimensional linear embedding layer to produce the original embedding sequence:

$$e \in \mathbb{R}^{N \times D_0} \tag{1}$$

Learnable positional embeddings of the same dimension are added to the original embedding sequence to utilize spatial prior information, resulting in the embedding:

$$Z^0 \in \mathbb{R}^{N \times D_0} \tag{2}$$

Z^0 serves as the input to the Transformer encoder, comprising L layers of Multi-Head Self-Attention (MSA). The self-attention (SA) mechanism, a core principle of the Transformer, updates the state of each embedding block by globally aggregating information at each layer:

$$SA(z_i) = softmax\left(\frac{q_i k^T}{\sqrt{D_h}}\right) v \tag{3}$$

Here, $[q, k, v] = zW_{qkv}$, where $W_{qkv} \in \mathbb{R}^{N\times 3D_h}$ is the projection matrix, and vectors $z_i \in \mathbb{R}^{1\times D_0}$, $q_i \in \mathbb{R}^{1\times D_h}$ are the *i*-th rows of z and q, respectively. MSA extends SA by concatenating multiple SAs and projecting the dimensions back to $\mathbb{R}^{D_0}$. Layer normalization is applied to the output of the last Transformer layer to obtain the encoded sequence:

$$Z^L \in \mathbb{R}^{N\times D_0} \tag{4}$$

Global features are acquired through n serially connected Transformer encoders. Considering training costs, this study sets n = 1.

After obtaining the feature maps from the two independent branches, the outputs from the CNN branch (local features) and the Transformer Encoder branch (global features) are merged. This creates a more informative composite feature map for further processing by the age recognition module. The model utilizes only two parallel networks without overly complex structures to effectively capture low-level spatial features and high-level semantic contexts, enhancing the model's inferential performance.

2.2 Improved spectrogram feature extraction algorithm

Spectrogram

A spectrogram is a three-dimensional spectrum, a graphical representation of how the speech spectrum changes over time, with frequency on the vertical axis and time on the horizontal axis. The intensity of a given frequency component at a specific time is indicated by the shade of gray or color tone at that point [15]. The construction process of a spectrogram is shown in Figure 3. Initially, the speech signal is converted into WAV format and preprocessed. Then, through frame segmentation and windowing operations, the speech signal is divided into several frames. In this paper, the Hamming window is used, and its expression is as follows:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] & 0 \le n \le N-1 \\ 0 & n = others \end{cases} \tag{5}$$

where *n* represents the frame index, and *N* represents the frame length.

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-jwm} \tag{6}$$

$$X_n(n, e^{jw}) = \sum_{m=0}^{N-1} x_n(m)e^{-jwm} \tag{7}$$

Next, the Short-Time Fourier Transform (STFT) and Discrete Fourier Transform (DFT) are applied to each frame of the speech signal *x*(*n*):

Let the frequency index be k, then the magnitude spectrum estimate $X_n(n,k)$ of the signal x(n) in the frequency domain is:

$$X_n(n,k) = \sum_{m=0}^{N-1} x_n(m)e^{-j\frac{2k\pi m}{N}} \tag{8}$$

The power spectral density function $P(n,k)$, also known as the power spectrum, is computed from the squared magnitude spectrum:

$$P(n,k) = |X(n,k)|^2 = \big(X(n,k)\big) \times \Big(conj\big(X(n,k)\big)\Big) \tag{9}$$

where $conj\big(X(n,k)\big)$ denotes the complex conjugate of $X(n,k)$.

To more intuitively represent the signal's energy, the power spectrum is converted to decibel (dB) units:

$$P_{dB}(n,k) = 10\lg\big(P(n,k)\big) \tag{10}$$

Finally, these energy values are organized into a matrix, and the matrix is converted into a two-dimensional image, thereby generating the spectrogram.

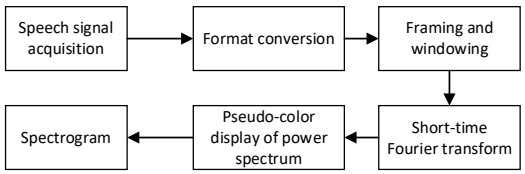

Fig. 3. Spectrogram Construction Process

Figure 4 displays a spectrogram of a segment of speech, where the light yellow areas represent voiced parts and the rest are unvoiced parts. The horizontal bars indicate the resonance peaks of the speech signal.

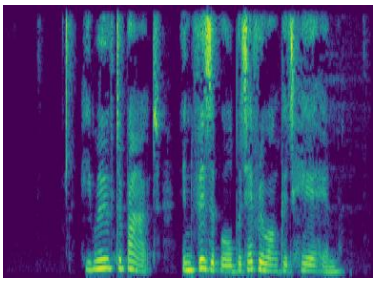

Fig. 4. Spectrogram of Speech

Spectrogram Feature Extraction Method Based on Frame Shift Strategy.

The conventional method of generating spectrograms typically segments speech signals into 1.5-second intervals, which can result in information distortion and thus lower recognition rates in systems. To maximally preserve the unique characteristics of speakers and increase the dataset size, this paper utilizes a frame shift strategy for spectrogram feature extraction. Each speaker's speech signal is finely segmented to generate multiple spectrograms.

Let the initial time be t_0, and the frame shift amount be Δt. Then, the starting time of the k-th segment is:

$$t_k = t_0 + k\Delta t \tag{11}$$

In this paper, initial segmentation is done over larger time intervals (e.g., 0-1.5s, 0.5-2.0s, 1-2.5s) until the end of the speech segment. More precise frame shifts are then performed within each larger time interval (e.g., 0-0.25s, 0.2-0.45s), with overlapping frames set to minimize information loss due to the Hamming window effect. Figure 5 illustrates some of the spectrograms generated using the frame shift strategy.

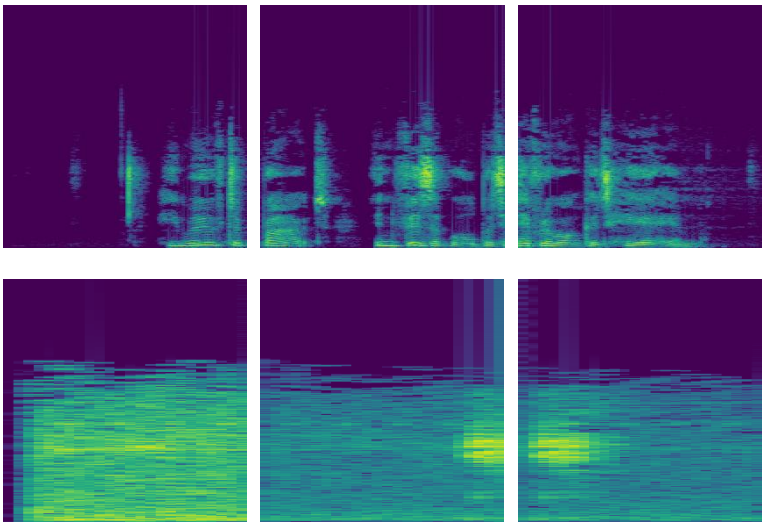

Fig. 5. Partial Spectrograms Generated Using Frame Shift Strategy

3 Experiment and Analysis

3.1 Databases

The Common Voice dataset [16] is currently the world's largest open-source speech dataset, used to develop optimized speech training technologies. It includes 32,123 hours of transcribed speech across 129 languages. The verified data also contain demographic metadata such as speaker age, gender, and accent. In this project, we use the English subset of the Common Voice dataset.

This paper utilizes 80 hours of data as the training sample and 1.5 hours of data for validation and testing. Table 1 shows the information of the speakers in the dataset, and Figure 6 displays the distribution of age and gender categories.

This paper utilizes 80 hours of data as the training sample and 1.5 hours of data for validation and testing. Table 1 shows the information of the speakers in the dataset, and Figure 6 displays the distribution of age and gender categories.

Table 1. Distribution of Age and Gender in the Common Voice Dataset

Age Groups	Train		Development		Test	
	Male	Female	Male	Female	Male	Female
Teens	4249	1060	84	28	82	34
Twenties	18494	3955	389	87	376	80
Thirties	13662	4560	256	88	295	92
Forties	8684	2187	180	60	184	47
Fifties	4946	4467	116	87	116	89
Sixties	2835	1722	55	40	43	44
Total	70821		1470		1482	

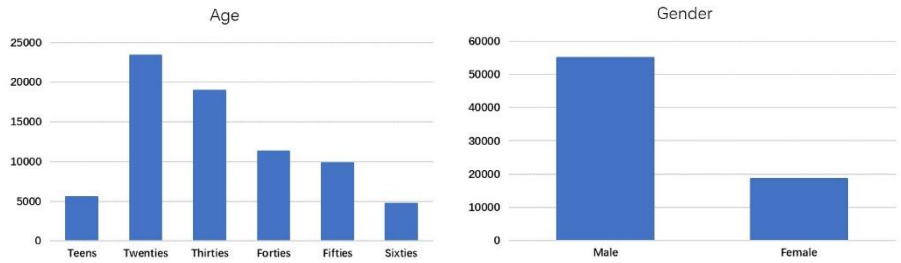

Fig. 6. Distribution of Age and Gender in the Common Voice Dataset

From Table 1 and Figure 6, it is evident that both age and gender distributions in the dataset are imbalanced. Data for males significantly outnumber that for females, and the data for adults in their twenties, thirties, and forties are substantially more prevalent than those for teenagers and older age groups (fifties, sixties).

3.2 Parameter Settings.

The experimental environment is on a Windows 11 operating system with an NVIDIA RTX 4080 SUPER graphics card. The model framework is implemented in TensorFlow [17], with the Adam optimizer, a learning rate of 0.0001, dropout set to 0.5, and iteration number set to 40. The model with the highest accuracy is retained for testing. The model evaluation metric used is the average recognition accuracy rate A, defined as follows:

$$A = \frac{\sum_{i=1}^{n} 1\ (P_i = T_i)}{n} \tag{12}$$

where P_i is the predicted label for the i^{th} instance, T_i is the true label for the i^{th} instance, $1\ (P_i = T_i)$ indicates that the value is 1 when $P_i = T_i$, and 0 otherwise, and n is the total number of samples.

3.3 Speaker age recognition based on CTPF-Net

Acoustic features may vary between genders [18]. To investigate whether the model proposed in this study can extract age-related acoustic features associated with gender, both gender-independent and gender-specific speech age recognition were conducted. The comparative results are presented in Table 2.

Table 2. Comparison of the effects of gender on age identification

Model Type	Average Recognition Accuracy Rate A%
Age	78.4
Age-Gender	84.5

According to Table 2, the gender-specific speech age recognition results are 6.1% higher than the gender-independent results. This indicates that gender information aids in distinguishing speech signals from different age groups, and the model proposed in this study is capable of learning age-related acoustic features associated with gender.

3.4 Result Analysis

To further validate the effectiveness of the model proposed in this paper, comparative experiments were conducted using the methods proposed in references [5], [9], [10], and [12]. The comparison results are shown in Table 3.

From Table 3, it can be deduced that the model presented in this paper achieves an average recognition accuracy of 84.5% on the Common Voice dataset, which represents an improvement compared to other methods.

Table 3. Comparison of Average Recognition Accuracy Rate (A) for Different Algorithms

	Input Feature	Average Recognition Accuracy Rate A%
[5]	WPMFC	63.2
[9]	x-Vector	74.8
[10]	Spectrogram	78.3
[12]	Spectro-temporal	80.2
CTPF-Net (Proposed)	Spectrogram	84.5

Figure 7 presents the confusion matrix for speech age recognition on the test set, showing that the model achieves accuracy rates of 92.1% and 90.9% for the female-fifties and female-sixties categories, respectively. It also performs well in the categories of female-twenties, female-thirties, female-forties; and male-teens, male-twenties, male-fifties, male-sixties. However, there are instances where some female-teens are misclassified as female-thirties, some male-thirties as male-twenties, and some male-forties as either male-twenties or male-thirties. This misclassification occurs because acoustic features such as pitch, fundamental frequency, and resonance peaks are similar between adjacent age groups. Additionally, due to the imbalanced data distribution,

smaller categories might experience overfitting, leading to a decrease in model performance.

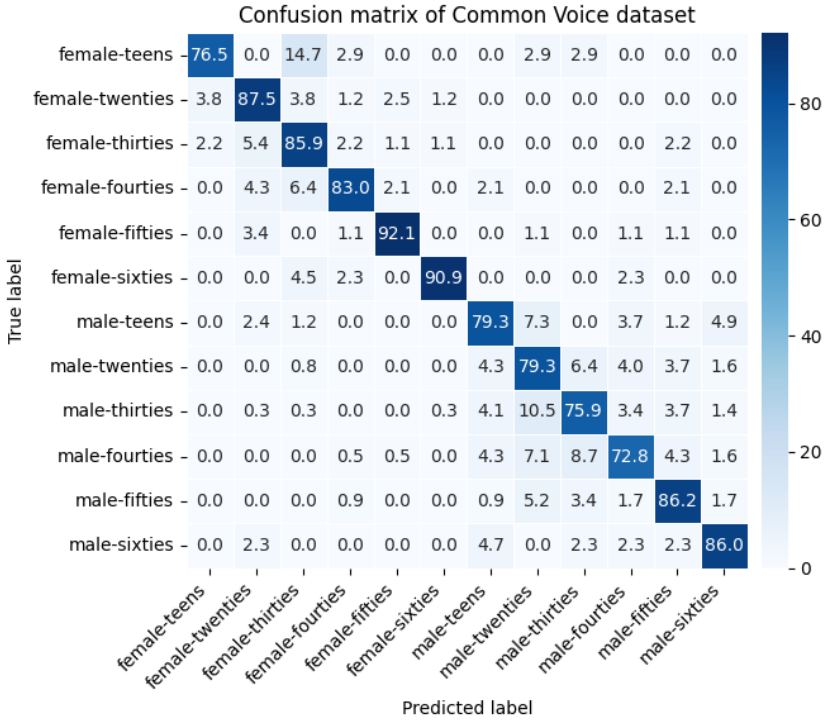

Fig. 7. Confusion Matrix for Speech Age Recognition

4 Conclusion

Recognizing attributes such as age and gender from speech signals enables more convenient and intelligent human-computer interactions. However, due to the complexity of speech signals and the variability and non-linearity of acoustic features, speech age recognition remains a challenging task. To address this issue, this paper draws on popular image processing methods and combines convolutional neural networks with Transformer architecture to propose a CNN and Transformer dual-branch parallel fusion network (CTPF-Net) for multidimensional feature extraction from spectrograms. This is followed by feature fusion and age classification using a classifier. To better preserve and extract information from spectrograms, a frame shift strategy for spectrogram feature extraction is employed, which preprocesses the original audio data and also compensates for issues such as the size and uneven distribution of the dataset.

To validate the performance of the proposed speech age recognition method, experiments were conducted on the English subset of the Common Voice dataset. The study first examined the impact of gender on speech age recognition and then compared the

model with acoustic feature methods and deep learning approaches. The results demonstrate that the proposed method effectively extracts gender-related speech age features and achieves an average accuracy of 84.5% in speech age recognition tasks.

While the method proposed in this paper shows an improvement in accuracy over other methods, it performs moderately in age classification for specific groups (such as female teens, male thirties, and male forties). Future work will focus on several areas: first, finding more significant speech age features and more efficient speech age recognition algorithms to reduce recognition errors between adjacent age groups; second, given that real-life acoustic environments are complicated by the presence of noise, more advanced models or methods (such as multimodal fusion) will be needed to develop more robust models. Finally, addressing performance inconsistencies due to imbalanced training data is another area worth exploring.

References

1. Ramakrishnan S, El Emary I M M. Speech emotion recognition approaches in human computer interaction[J]. Telecommunication Systems, 2013, 52: 1467-1478.
2. Hanifa R M, Isa K, Mohamad S. A review on speaker recognition: Technology and challenges[J]. Computers & Electrical Engineering, 2021, 90: 107005.
3. Yue M, Chen L, Zhang J, et al. Speaker age recognition based on isolated words by using SVM[C]//2014 IEEE 3rd International conference on cloud computing and intelligence systems. IEEE, 2014: 282-286.
4. Chen O T C, Gu J J. Improved gender/age recognition system using arousal-selection and feature-selection schemes[C]//2015 IEEE International Conference on Digital Signal Processing (DSP). IEEE, 2015: 148-152.
5. Du X N, Yu Y B. Multi Resolution Feature Extraction of Effective Frequency Bands for Age Recognition[J]. Journal of Signal Processing, 2016, 32 (09): 1101-1107.
6. Bahari M H, McLaren M, van Leeuwen D A. Speaker age estimation using i-vectors[J]. Engineering Applications of Artificial Intelligence, 2014, 34: 99-108.
7. Lortie C L, Thibeault M, Guitton M J, et al. Effects of age on the amplitude, frequency and perceived quality of voice[J]. Age, 2015, 37: 1-24.
8. Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
9. Ghahremani P, Nidadavolu P S, Chen N, et al. End-to-end Deep Neural Network Age Estimation[C]//Interspeech. 2018, 2018: 277-281.
10. Tursunov A, Mustaqeem, Choeh J Y, et al. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms[J]. Sensors, 2021, 21(17): 5892.
11. Sánchez-Hevia H A, Gil-Pita R, Utrilla-Manso M, et al. Age group classification and gender recognition from speech with temporal convolutional neural networks[J]. Multimedia Tools and Applications, 2022, 81(3): 3535-3552.
12. Mavaddati S. Voice-based age, gender, and language recognition based on ResNet deep model and transfer learning in spectro-temporal domain[J]. Neurocomputing, 2024, 580: 127429.
13. Akan A, Cura O K. Time–frequency signal processing: Today and future[J]. Digital Signal Processing, 2021, 119: 103216.

14. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.
15. Anusuya M A, Katti S K. Front end analysis of speech recognition: a review[J]. International Journal of Speech Technology, 2011, 14: 99-145.
16. Ardila R, Branson M, Davis K, et al. Common voice: A massively-multilingual speech corpus[J]. arXiv preprint arXiv:1912.06670, 2019.
17. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems[J]. arXiv preprint arXiv:1603.04467, 2016.
18. Torre III P, Barlow J A. Age-related changes in acoustic characteristics of adult speech[J]. Journal of communication disorders, 2009, 42(5): 324-333.