Proceedings of the 43rd Chinese Control Conference
July 28-31, 2024, Kunming, China

# A long-duration Speech Semantic Recognition and Summarization Model for multi-speaker Conversations

Mingzhen Song[1], Weifeng Liu*[1], Chuanzhao Xu[2], Jiating Wang[1]
1. Shaanxi University of Science and Technology, Xi'an 710000, P. R. China
E-mail: 13649372256@163.com
2. China Jikan Research Institute of Engineering Investigations and Design, Co, Ltd, Xi'an 710016, P. R. China
E-mail: 545039451@qq.com

**Abstract:** This paper aims to propose a comprehensive solution to address challenges in the field of audio processing for long-duration, multi-speaker conversational meetings. The identified issues include multi-speaker voice separation, difficulties in handling long temporal sequences, and the high redundancy in generating speech scene text. The proposed solution encompasses four modules: Firstly, addressing the issue of complex background noise interference in speech signals by introducing Deep Extractor for Music Sources with extra Components (Demucs) audio separation and denoising technology to effectively extract and suppress noise. Secondly, employing a combined approach of Convolutional Neural Network (CNN) and self-attention to extract multi-scale features for speaker identification in a multi-speaker environment. Thirdly, the Conformer model is trained using a large-scale Chinese speech database to tackle challenges in processing extended temporal sequences, ensuring the accurate transformation of speech information into textual representations. Finally, the latest Generative Pre-trained Transformer-3.5 (GPT-3.5) model undergoes fine-tuning using a Chinese text summarization dataset. This process enables the generation of text summaries, extracting key points from multi-speaker long-duration conversations. The effectiveness of the proposed algorithms is validated through speech scene experiments in various scenarios.

**Key Words:** Speech Recognition, Voiceprint Recognition, Text Summarization

## 1 Introduction

Text generation technology in the context of speech has a wide range of applications. For example, it is used in automatic subtitle generation, intelligent assistants, voice search, and automated report generation. Furthermore, with the increasing popularity of video conferencing, there is a growing demand for the recording and subsequent organization and analysis of long-duration audio. This requires extracting key text information from lengthy audio content to provide a better user experience.

In the context of long-duration speech scenarios, text generation technology primarily faces challenges in noise handling by speech enhancement techniques, context-aware long-duration speech recognition, and maintaining context consistency in natural language processing techniques, as well as issues related to precision extraction.

To address the aforementioned issues, this paper combines the strengths of CNN and Transformer technologies and proposes a method for conference speech enhancement in long-duration multi-speaker conversations, coupled with text summarization using GPT-3.5 technology. The approach is divided into four modules, which encompass speech enhancement, speaker recognition, long-duration speech recognition, and text summarization. Firstly, to tackle the challenge of complex background noise interference, effective speech signal extraction and noise suppression are achieved by introducing Demucs [1] audio separation and denoising techniques. Secondly, a combined approach utilizing CNN and self-attention is employed to extract multi-scale features of speakers and identify multiple speakers, integrating personal voiceprint models. Thirdly, Voice Activity Detection (VAD) technology and the Conformer model are used for long-sequence speech recognition to address challenges in processing extended audio sequences, accurately converting speech information into textual representations. Lastly, fine-tuning is performed using the natural language processing (NLP) model GPT-3.5 for automatic text summarization generation. This approach aims to enhance the accuracy and practicality of text generation in speech scenarios.

The structure of this paper is as follows: Chapter 2 presents the current research status, Chapter 3 introduces the problem background, Chapter 4 describes the algorithms, Chapter 5 presents the analysis of experimental results and performance evaluations, and the final section is the conclusion of this paper.

## 2 Research Status

Currently, deep learning has made significant progress in addressing challenges related to speech enhancement, speech recognition, and natural language processing [2–4]. However, challenges persist in the recognition and understanding of long-duration speech in complex environments. To address this, a comprehensive review is conducted focusing on speech recognition and text understanding.

In recent years, the rise of deep learning technologies has brought about new breakthroughs in speech recognition. Its applications in speech recognition mainly encompass models such as Deep Convolutional Neural Networks (DCNN) [5], Long Short-Term Memory networks (LSTM)[6], and Transformer[7]. These models are capable of automatically learning representations of input features and undergoing training on large-scale data, greatly enhancing the accuracy and performance of speech recognition. End-to-end speech recognition technology simplifies the training and inference process by integrating acoustic and language models into a unified network. Techniques such as transfer learning and re-

*Corresponding author. This work was supported by the NSFC under Grant 62376147, and Natural Science Basic Research Program of Shaanxi (2022JQ-601).

inforcement learning have also been introduced into speech recognition to improve the model's generalization ability in specific tasks and scenarios[8, 9].

Text understanding is one of the core challenges in natural language processing. Machine learning, especially deep learning, plays a crucial role in text summarization. Models based on deep learning typically employ structures like RNN, LSTM, Transformer, etc., to encode input text and generate summaries through a decoder. In recent years, with the introduction of pre-trained language models such as BERT [10] and GPT [11], leading to significant improvements in text summarization. Presently, multimodal text generation has emerged as a novel information processing approach. By integrating different types of data, such as text [12], images[13], audio[14], etc., it offers a more comprehensive and in-depth summarization. Currently, multi-modal text generation based on deep learning has been widely used in various fields, such as text mining [15], semantic analysis[16], medical diagnosis [17], etc.

## 3 Background and problem description

Demucs consists of a multi-layer convolutional encoder and decoder, featuring U-net[12] skip connections and applying a sequence modeling network on the encoder's output. It is characterized by its number of layers $L$, initial hidden channel count $H$, layer kernel size $K$, stride $S$, and upsampling factor $U$. The encoder and decoder layers are numbered from 1 to $L$ (in reverse order for the decoder, so layers at the same scale have the same index). The model has a single-channel input and output.

Formally, given an audio signal $x \in \Re^T$, it is composed of the superposition of the real speech signal $y \in \Re^T$ and the additional background signal $n \in \Re^T$ , that is $x = y + n$, with the length $T$ representing the duration of the signal, which is not a fixed value.

Encoder network $E_n$ takes an audio signal $x$ as input and outputs its latent representation $E_n(x) = z$. Each layer consists of a convolutional layer with a kernel size of $K$ and a stride of $S$, with $2^{i-1}H$ output channels, followed by a ReLU activation, a 1x1 convolution with $2^iH$ output channels, and finally a GLU activation that converts the channel count back to $2^{i-1}H$.

Next, a sequence modeling network $R$ takes the latent representation $z$ as input and outputs a nonlinear transformation of the same size, denoted as $R(z) = LSTM(z) + z$, represented as $\hat{z}$. The bidirectional LSTM network consists of 2 layers and $2^{L-1}H$ hidden units. This is followed by a linear layer used to merge the two outputs.

Finally, the decoder network $D$ takes $\hat{z}$ as input and outputs an estimate $D(\hat{z}) = \hat{y}$ of the speech signal. The $i$-th layer of the decoder takes $2^{i-1}H$ channels as input and applies a 1x1 convolution with $2^iH$ channels, followed by a GLU activation, which outputs $2^{i-1}H$ channels. Then, a transposed convolution with a kernel size of 8 and a stride of 2 is applied, resulting in $2^{i-2}H$ output channels, accompanied by a ReLU function. For the last layer, the output is a single channel without ReLU. A skip connection connects the output of the $i$th layer of the encoder to the input of the ith layer of the decoder, as shown in Fig. 1. The arrows represent the skip connections of the UNet.

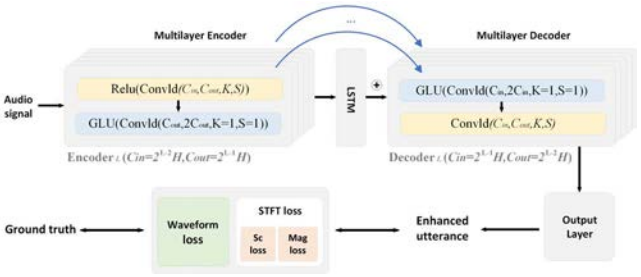


Fig. 1: Demucs Structural Schematic

## 4 Algorithm in This Paper

### 4.1 Speech Signal Enhancement

For any audio signal with an input length of $T$, denoted as $y$, its estimate is given as $\hat{y}$. The loss function minimized includes $L_1$ loss on the waveform (in the time domain) and spectral loss using multi-resolution Short-Time Fourier Transform (STFT) (in the time-frequency domain), as shown in the formula (1):

$$L_{DEMUCS}(y, \hat{y}) = \frac{1}{T}\|y - \hat{y}\|_1 + \sum_{i=1}^{M} L_{stft}^{(i)}(y, \hat{y}) \quad (1)$$

Where $M$ represents the number of STFT losses, each $L_{stft}^{(i)}$ is applied to STFT losses at different resolutions, and $\|\|_1$ denotes the $L_1$ norm operation. The superscript $i$ indicates an index corresponding to different frame sizes and frame shifts.

Furthermore, certain index $(i)$ terms of $L_{stft}(y, \hat{y})$ consist of two components: spectral convergence (indicated by the subscript sc) and spectral magnitude (indicated by the subscript mag), as presented in the formulas (2), (3) and (4):

$$L_{stft}(y, \hat{y}) = L_{sc}(y, \hat{y}) + L_{mag}(y, \hat{y}) \quad (2)$$

$$L_{sc}(y, \hat{y}) = \frac{\|\,|STFT(y)| - |STFT(\hat{y})|\,\|_F}{\|\,|STFT(y)|\,\|_F} \quad (3)$$

$$L_{mag}(y, \hat{y}) = \frac{1}{T}\|\log|STFT(y)| - \log|STFT(\hat{y})|\|_1 \quad (4)$$

Where $\|\cdot\|_F$ and $\|\cdot\|_1$ represent the Frobenius norm and $L_1$ norm, respectively. The multi-resolution STFT loss is defined as the sum of all STFT loss functions using different STFT parameters.

### 4.2 Speaker Verification

ERes2Net incorporates local feature fusion (LFF) and global feature fusion mechanism (GFF) [18] into the traditional Res2Net network. It combines both local and global feature fusion methods to improve recognition performance. The specific structure see Fig. 2.

LFF component aims to enhance speaker embeddings' distinctiveness at a finer level (see Fig. 3). It utilizes the Attention Feature Fusion (AFF) module to aggregate neighboring feature maps within residual blocks (see Fig. 4). GFF component adjusts features at different time scales, enhancing the robustness of speaker embeddings from a global perspective.

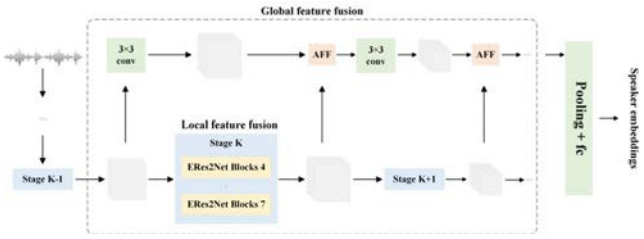


Fig. 2: The structure of the ERes2Net model

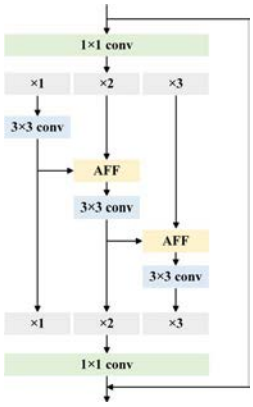


Fig. 3: LFF

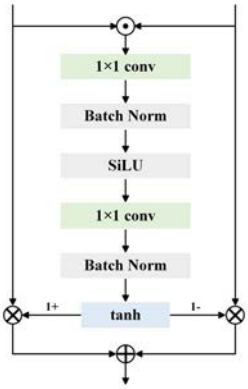


Fig. 4: AFF

The output of the ERes2Net block is shown in formula (5).

$$y_i = \begin{cases} K_i(x_i), i = 1 \\ K_i(U(x_i, y_{i-1}) + 1) \cdot x_i \\ +(1 - U(x_i, y_{i-1})) \cdot y_{i-1}, i > 1 \end{cases} \quad (5)$$

Where $U(\cdot)$ represents the AFF module, which is used to calculate the local attention weights for adjacent feature maps. The AFF module takes the concatenation of neighboring feature maps $x_i$ and $y_{i-1}$ as its input. Then, the local attention weights U are computed according to formula (6).

$$U = \tanh(BN(W_2 \cdot SiLU(BN(W_1 \cdot [x_i, y_{i-1}])))) \quad (6)$$

Here, $[\cdot]$ represents concatenation along the channel dimension. $W1$ and $W2$ are point-wise convolutions with output channel sizes of $C/r$ and $C$, respectively. $r$ is the channel reduction ratio (set to $r = 4$ in this paper). BN stands for batch normalization. $SiLU(\cdot)$ and $tanh(\cdot)$ represent the Sigmoid Linear Unit (SiLU) and hyperbolic tangent activation functions, respectively. This module aims to dynamically weight and combine features based on their importance, enhancing the model's ability to extract relevant information from the input signal.

GFF component of ERes2Net (see Fig.2) selects the multi-scale feature $\{S_j, j = 2, 3, 4\}$ from the last layer of each ERes2Net stage. It uses a $3 \times 3$ convolution kernel to downsample the higher-resolution feature maps in both the time and frequency dimensions, doubling the channel dimension. The outputs of each stage are then concatenated, and the bottom-top channel attention is calculated through the AFF module to adjust $\{S_j, j = 2, 3, 4\}$. The AFF module calculates attention weights from a global perspective. The downsampled feature maps are enhanced by the AFF module (see formula (7)).

$$F_j = U[D(S_{j-1}), S_j], j = 2, 3, 4 \quad (7)$$

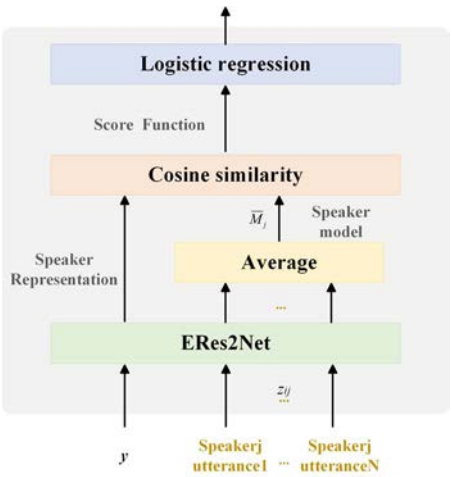


Fig. 5: Speaker models averaged from the voiceprint library

Here, $D(\cdot)$ represents the downsampling operation. $F_j$ represents the fusion of the outputs from the bottom-top channels across the $(j-1)$ stages and the output of the $j$-th stage.

To enhance the system's robustness against different environments and speaker variations, improve the stability of speech features, this paper builds upon ERes2Net by adding a speaker enrollment library. In this library, audio features from the same person are averaged to create individual speaker models. These speaker models are then compared for similarity with new audio features (see Fig. 5).

Using $z_i$ to represent the feature vector of the $i$-th audio sample, for multiple audio samples from the same speaker, their feature vectors are averaged to create an individual speaker model (Speaker model) $M_j$, where $j$ represents the speaker's identifier. The vector representation of the individual speaker model is denoted as $\bar{M}_j$ (see formula (8)).

$$\bar{M}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} z_{ij} \quad (8)$$

Where $N_j$ represents the number of audio samples for speaker $j$, and $z_{ij}$ is the feature vector of the $i$-th sample for speaker $j$. For the feature vector $y$ of a new audio sample, its similarity with the individual speaker model $M_j$ is calculated using cosine similarity (see formula (9)).

$$(y, \bar{M}_j) = \cos(\theta) = \frac{y \cdot \bar{M}_j}{\|y\| \left\|\bar{M}_j\right\|} \quad (9)$$

Where $\cdot$ represents the dot product of vectors, and $\|\cdot\|$ represents the norm of the vectors. By comparing the similarity values, it is possible to determine the similarity between the new audio sample and the individual speaker model, enabling speaker identification or speaker verification.

### 4.3 Speech Recognition

This paper employs Conformer end-to-end models for speech recognition, combining Transformer models and Convolutional Neural Networks (CNN), utilizes the Transformer to model the speech features in the input sequence and leverages self-attention mechanisms to capture dependencies between different positions [19]. The CNN network

within the model is used to extract local features from the input sequence, assisting the model in capturing local patterns and structural information within the input sequence. The Conformer module is depicted Fig. 6.

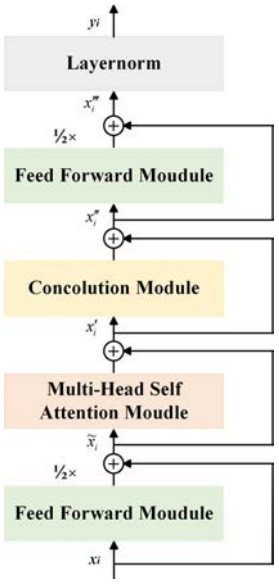


Fig. 6: Conformer Block[19]

In the Convolution Module, Pre-norm residual, point-wise convolution, and Gated Linear Unit (GLU) are employed. This is illustrated in Fig. 7.

The input $x_i$ to Conformer module $i$ and its output $y_i$ can be represented by formula (10):

$$\begin{aligned} \tilde{x}_i &= x_i + \tfrac{1}{2}FFN(x_i) \\ x'_i &= \tilde{x}_i + MHSA(\tilde{x}_i) \\ x''_i &= x'_i + Conv(x'_i) \\ y_i &= Layernorm(x''_i + \tfrac{1}{2}FFN(x''_i)) \end{aligned} \tag{10}$$

Where FFN, MHSA, Conv, and LayerNorm represent the Feed-Forward Network module, Multi-Head Self-Attention module, Convolution module, and Layer Normalization module, respectively. $\tilde{x}_i$, $x'_i$, $x''_i$, and $x'''_i$ correspond to the outputs of different stages within the Conformer module as illustrated in Fig. 6.

This article presents the overall model architecture as shown in Fig. 8.

## 5 Experimental Analysis

### 5.1 Dataset

The training dataset used in the speaker recognition experiments in this paper is the CN-Celeb ultra-large dataset, which includes data from over 200,000 Chinese speakers. The test dataset is the CN-Celeb-test dataset.

For the speech recognition experiments, the training dataset consists of the WenetSpeech training dataset (10,000+ hours) and additional Chinese speech data (3,000+ hours). The test dataset includes WenetSpeech test data,


Fig. 7: Convolution Module

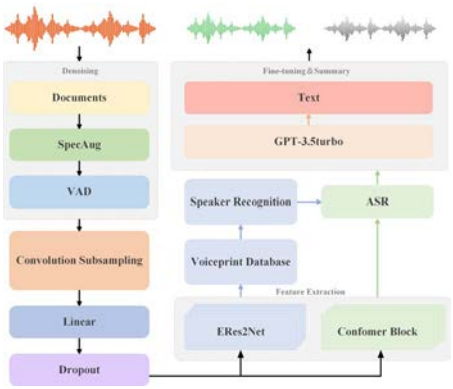


Fig. 8: The diagram of the overall model architecture in this article

which is divided into test-net and test-meeting sections, as well as the Aishell Chinese Mandarin speech dataset.

In the fine-tuning experiments with GPT3.5-turbo, the LCSTS large-scale Chinese short-text summarization dataset is used. This dataset is sourced from Sina Weibo and contains 2.4 million records. 10,000 records are extracted as the test set, while the remaining data is used for training. The data is transformed into a format compatible with the GPT3.5-turbo model.

### 5.2 Experimental Setup

The experiments were conducted on a server equipped with an NVIDIA RTX A6000 GPU, 79.4GB of RAM, and 1TB SSD storage. The server ran on an Intel(R) Core(TM) i9-12900K processor and used the Windows 11 operating system.

For the speech signal enhancement experiments, the following hyperparameters were used: $U$=2, $S$=2, $K$=8, $L$=5, $H$=64. Before feeding the input to the model, input normalization was performed based on its standard deviation, and the output was scaled back using the same factor.

For the speaker recognition experiments, the ResNet[18] served as the backbone network. The AFF self-attention mechanism was incorporated to extract both global and local features from the speech and fuse them. The AAM-loss function was used with a parameter count of 55.165024 M.

In the speech recognition experiments, the Conformer model was used as the main network architecture with 8 attention heads. During training, the CTC loss was minimized using backpropagation.

In the text summarization experiments, we utilized the fine-tuning feature provided by OpenAI, with GPT-3.5-turbo as the base model, employing the Adam optimizer with an initial learning rate set to 5e-5 and a batch size of 16. The model underwent 4 training epochs on the training set.

### 5.3 Evaluation Metrics

In speaker recognition, a threshold is used to determine whether to accept or reject authentication from the speaker recognition system. The evaluation metrics for speaker recognition include Equal Error Rate (EER) and the Minimum Detection Cost Function (MinDCF). EER refers to the error rate at the point where the True Positive Rate (TPR) equals the False Positive Rate (FPR). It represents the point

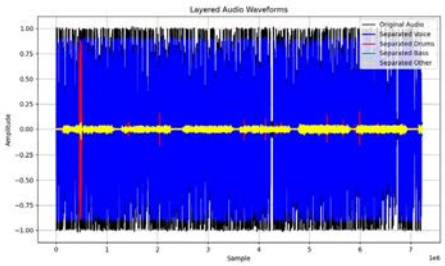


Fig. 9: Comparison of waveform plots

at which the recognition system can correctly identify or incorrectly reject with equal probability. EER is commonly used to measure system performance and is shown in formula (11).

$$EER = \frac{FPR + FNR}{2} \tag{11}$$

Where FNR stands for False Negative Rate.

MinDCF is used to assess the overall performance of the speaker recognition system. It takes into account both the false acceptance rate and false rejection rate while considering the cost differences associated with different types of errors. MinDCF can be expressed as shown in formula (12).

$$\begin{aligned} MinDCF = \min(&C_{miss} \cdot P_{miss} \cdot P_{target} \\ &+C_{fa} \cdot P_{fa} \cdot (1 - P_{target})) \end{aligned} \tag{12}$$

Where $C_{miss}$ represents the cost associated with missed detections (Miss Cost), $C_{fa}$ is the cost associated with false alarms (False Alarm Cost), $P_{miss}$ is the FNR, $P_{fa}$ is the FPR, $P_{target}$ is the Prior Probability, which represents the prior probability of genuine speakers and impostors.

In the context of speech recognition, Word Error Rate (WER) is a commonly used evaluation metric to measure the degree of dissimilarity between the recognized text and the reference text. WER calculation involves four values: Insertions ($I$), Deletions ($D$), Substitutions ($S$), and Total Number of Words ($N$), as shown in formula (13):

$$WER = \frac{S + D + I}{N} \tag{13}$$

### 5.4 Experimental Results

Model evaluation is conducted on two subtasks: speaker recognition and speech recognition.

In the speech signal enhancement experiment, the results are shown in Fig. 9.

In Fig. 9, Black represents the original audio, and blue corresponds to the human voice part.

The enhanced experimental results see Table 1 and Table 2.

Table 1: Voiceprint Recognition Evaluation

| metric / model | threshold | EER | MinDCF |
|---|---|---|---|
| | CN-Celeb-test | | |
| CAM++ | 0.29 | 0.04765 | 0.31436 |
| Res2Net | 0.29 | 0.14875 | 0.58452 |
| Deepspeech2 | / | / | / |
| Squeezeformer | / | / | / |
| OURS | 0.36 | 0.03541 | 0.20317 |

Table 2: Speech Recognition Evaluation

| metric / model | WER | | |
|---|---|---|---|
| | aishell_test | test_net | test_meeting |
| CAM++ | / | / | / |
| Res2Net | / | / | / |
| Deepspeech2 | 0.05478 | 0.13457 | 0.20501 |
| Squeezeformer | 0.03129 | 0.11021 | 0.19142 |
| OURS | 0.03114 | 0.11732 | 0.18346 |

Tables 1 and 2 present the comparative results of the proposed method with CAM++, Res2Net, Deepspeech2, and Squeezeformer on the test datasets CN-Celeb-test, aishell-test, test-net, and test-meeting. The speaker recognition model CAM++ employs a backbone based on a Dense Time-Delay Neural Network (D-TDNN). DeepSpeech 2, released by Baidu, is an end-to-end speech recognition system primarily based on Deep RNNs or LSTMs. Squeezeformer replaces the standard Conformer's sequence of ffn + self-attention + conv module + ffn with a combination of self-attention + ffn + conv module + ffn (MFCF). As shown in the tables, for the task of voiceprint recognition, even with a higher threshold than other models (imposing stricter requirements for system authentication), the proposed algorithm still shows superior performance, with the error rate EER approximately 0.012 or about 25.5% lower than CAM++, and more than 0.11 lower than Res2Net, significantly improving accuracy; the overall system cost MinDCF is reduced by about 0.11 compared to CAM++, and approximately 0.38 compared to Res2Net. In the task of speech recognition, compared to Deepspeech2, the proposed model reduces the word error rate by 42.59%, 15.3%, and 10.00% on the three different types of test datasets, respectively; the difference is marginal compared to the Squeezeformer model, with slightly better performance on individual datasets. The proposed model exhibits a good overall performance in both evaluation tasks.

The visual example of the model in this article is shown in Fig. 10.

## 6 Conclusion and Outlook

This paper combines the advantages of CNN and Transformer technologies, proposing a method for long-duration speech multi-speaker dialogue conference speech enhancement and text summary generation based on GPT-3.5 technology. The process is divided into four modules: speech enhancement, speaker identification, long-duration speech content recognition, and text summary generation. Firstly, Demucs audio separation and noise reduction technology is introduced to suppress noise in the speech signal. Secondly, a method combining CNN and self-attention is used to extract multi-scale features of the speaker, combined with an individual voiceprint model to identify multiple speakers' identities. Thirdly, VAD technology and the Conformer model are utilized for long-duration speech recognition to address the challenges of long-time sequence processing and accurately convert speech information into text representation. Finally, the natural language processing NLP model GPT-3.5 is fine-tuned for automatic text summary generation.

However, the current model is primarily designed for of-


Fig. 10: Visual Example of The Model

fline speech processing, and its application in real-time processing systems is limited. This is mainly because the demands of processing long-time sequences and the requirements for high real-time performance pose greater challenges to the model's response speed and processing capabilities. In recent developments, it is worth considering structural optimizations of the Efficient Transformer model, which has a stronger capability for processing long sequences, as well as applications of model compression techniques (such as distillation and quantization) and hardware acceleration, to overcome the current system's limitations in real-time speech processing.

## References


[1] Defossez A, Usunier N, et al.,Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254, 2019*.

[2] Xu X, Tu W, Yang Y,CASE-Net: Integrating local and non-local attention operations for speech enhancement. *Speech Communication*. 2023, 148: 31-39. ISSN 0167-6393.

[3] Radford A, Xu T, et al., Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning*. PMLR, 2023: 28492-28518.

[4] Chen S, Ramirez N, et al., Deep Learning-Based Natural Language Processing to Automate Esophagitis Severity Grading from the Electronic Health Records. *International Journal of Radiation Oncology, Biology, Physics*. 2023, 117(2): S18.

[5] Zhu Y, Zeng Q, Continuous Speech Recognition Based on DCNN-LSTM.*2023 5th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP)*,Chengdu, China, 2023: 1247-1250.

[6] Hochreiter S, Schmidhuber J, Long short-term memory.*Neural Computation*.1997, 9(8): 1735-1780.

[7] Vaswani A, Shazeer N,et al.,Attention is all you need.in*Advances in neural information processing systems*.2017: 5998-6008.

[8] Zhen-Tao L ,Bao-Han W ,et al.,Speech emotion recognition based on meta-transfer learning with domain adaption.*Applied Soft Computing*.2023:147

[9] Lai V D, Tan H, et al., Boosting Punctuation Restoration with Data Generation and Reinforcement Learning.*arXiv preprint arXiv:2307.12949, 2023*.

[10] Devlin J, Lee K, Toutanova K,BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,Volume 1 (Long and Short Papers), 2018: 4171-4186.

[11] Radford A, Narasimhan K, et al.,Improving Language Understanding by Generative Pre-training.*URL:https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language-understanding-paper.pdf*, 2018.

[12] He, B, Wang, J,et al.,Align and Attend: Multimodal Summarization With Dual Contrastive Losses.in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,2023: 14867-14878.

[13] Ding N, Deng C,et al.,Image Captioning With Controllable and Adaptive Length Levels.*Advances in neural information processing systems IEEE Transactions on Pattern Analysis and Machine Intelligence*.2023,DOI: 10.1109/TPAMI.2023.3328298.

[14] Wilschut T, Sense F, van Rijn H,Speaking to remember: Model-based adaptive vocabulary learning using automatic speech recognition.*Computer Speech and Language*.2024, Volume 84, 101578. ISSN 0885-2308.

[15] Ren X, Li Y, Guo M,Dynamically Identifying and Evaluating Key Barriers to Promoting Prefabricated Buildings: Text Mining Approach.*Journal of Construction Engineering and Management*.2023, 149(9): 04023075.

[16] Malik K, Widyarini I G A A, et al.,Differences in syntactic and semantic analysis based on machine learning algorithms in prodromal psychosis and normal adolescents.*Asian Journal of Psychiatry*.2023, 85: 103633.

[17] Azam, K. B, et al.,A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics.*Computers in biology and medicine*.2022,144, 105253.

[18] Chen Y, Wang H, et al., An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. *arXiv preprint arXiv:2305.12838, 2023*.

[19] Gulati A, Qin J, et al., Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100, 2020*.