Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 13
AMMSAC 2026

Research on Multi-voice Discourse Structure Analysis and Information Purification Based on Neural Network Technology

Caizhi Fei [1, *], Xiaoqing Liu [2, a] and Siyu Qi [3, b]

[1] School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang 330000, China

[2] School of International Economics and Politics, Jiangxi University of Finance and Economics, Nanchang 330000, China

[3] School of Statistics and Data Science, Jiangxi University of Finance and Economics, Nanchang 330000, China

* Corresponding Author Email: 2202402003@stu.jxufe.edu.cn, [a] 2202405299@stu.jxufe.edu.cn, [b] 2202405297@stu.jxufe.edu.cn

Abstract. The overlapping and aliasing of multi-voice signals in complex scenes can easily lead to the blurring of discourse boundaries, the disorder of hierarchical structure and the distortion of key semantic information, which is a key and difficult problem in the field of speech analysis and information processing. Based on the neural network model, this paper constructs a time series feature modeling mechanism for the mixed speech characteristics of multi-person voice, excavates the deep correlation features of speech signals, and accurately disassembles the hierarchical discourse structure of multi-person dialogue. At the same time, through the adaptive redundant interference suppression strategy, the background noise and cross-human voice interference are effectively filtered out, and the accurate purification of the core speech information is completed. The experimental results show that this method can effectively improve the performance of multi-voice analysis and purification in complex aliasing scenarios, and can provide reliable technical reference for related engineering applications such as intelligent speech analysis and speech noise reduction.

Keywords: Multi-voice analysis; neural network; discourse structure; feature modeling; information purification; speech signal processing.

1. Introduction

With the wide application of intelligent speech technology in scene monitoring, human-computer interaction, speech forensics and other fields, the demand for high-precision processing of multi-person aliasing speech in complex environments is increasingly urgent. The speech signals collected in the actual scene generally have problems such as human voice overlap, discourse interpenetration and background noise interference, which are easy to cause discourse structure disorder, boundary blurring and semantic information redundancy, and seriously restrict the processing accuracy and stability of downstream tasks such as human voice analysis, feature extraction and semantic analysis. It is the core bottleneck of the current complex speech intelligent processing technology.

The existing multi-person voice processing schemes mostly focus on voice separation or single noise reduction optimization. Traditional signal processing methods rely on artificial feature design, which is difficult to adapt to multi-person cross-talk scenarios and can not effectively describe the dialogue hierarchy. Although the mainstream deep learning methods have excellent deep feature extraction capabilities, they generally ignore the temporal correlation modeling of discourse structure and lack targeted adaptive interference suppression strategies, which can easily lead to the loss of key speech information and the distortion of analytical results. It is difficult to balance the accuracy of discourse structure analysis and the quality of speech information purification.

Based on this, this paper first sorts out the research status of multi-person voice processing and speech purification, and then details the core method system of time series modeling, structural analysis and

information purification. The performance and module effectiveness of the model are verified by multiple sets of comparison and ablation experiments. Finally, the research results are summarized and the follow-up optimization direction of complex voice intelligent processing technology is prospected.

2. Related Work

The rapid iteration of information technology and network technology has promoted the wide application of neural network models in many intelligent sensing fields such as image, text, and speech recognition, and has made breakthroughs. Krizhevsky et al. [1] used Convolutional Neural Network (CNN) to solve the problem of image classification for the first time in the 2012 ImageNet visual competition. The classification results are beyond the traditional image classification methods. This work proves the effectiveness of convolutional neural network in complex models, and then gradually establishes its dominant position in computer vision.

With the transformation of modern linguistic research from language form structure to functional mechanism, discourse analysis has developed into an independent key research direction, focusing on the dimensions of discourse structure, generation rules and expression patterns. The discourse has both dynamic generation process and hierarchical structure attributes, and each structural unit has independent evolution law and exclusive realization mechanism [2]. The analysis of dialogue structure and operation law is the core content of discourse analysis, which mainly covers the whole and local double-layer structure of dialogue, including turn change, dialogue unity and typical interaction phenomena such as discourse overlap, insertion, interruption and correction.

In order to meet the requirements of lightweight deployment of terminal equipment, various efficient and lightweight network architectures based on information distillation have been proposed one after another. Hui et al. [3] first proposed the information distillation network (IDN), which divides the previously extracted features into two parts along the channel dimension. One part is continued by the subsequent convolution layer, and the other part is retained and processed. The feature aggregation, IDN uses this channel segmentation method to achieve good performance on a moderate scale. Then Hui et al. [4] proposed the information multi-distillation network (IMDN). By designing a cascade information multi-distillation module, the hierarchical features are gradually extracted to further improve the reconstruction performance. Liu et al. [5] reconsidered the architecture of IMDN and proposed RFDN, which uses feature distillation connection instead of channel segmentation to realize multi-level progressive refinement distillation of features. On the basis of RFDN, Kong et al. [6] removed the feature distillation connection and proposed the residual local feature network (RLFN), which accelerated the inference speed of the network. Li et al. [7] improved the convolution operation and proposed a blueprint separable residual network (BSRN), which uses the blueprint separable convolution to reduce the redundant calculation in the feature extraction block. Combined with a more effective attention module, the extraction ability of distillation features is enhanced.

3. Model

3.1. Multi-vocal temporal feature modeling

In order to accurately capture the deep correlation information of multi-person mixed speech in complex scenes and solve the problem of feature coupling caused by human voice overlap and discourse interpenetration, this paper constructs a multi-person voice timing feature modeling module based on neural network to complete the feature mapping and dimension representation of the original mixed speech signal. The original input multi-voice mixed time-domain signal is $x(t)$, which contains multi-channel voice, background noise and cross-interference components. The basic feature extraction is completed by the neural network pre-feature extraction layer. The calculation method is:

$$F_{init} = \mathcal{F}(x(t); \theta)$$

where $\mathcal{F}(\cdot)$ is the mapping function of network feature extraction; θ is a network learnable parameter; F_{init} is the initial speech feature map. On this basis, according to the temporal continuity characteristics of speech, the temporal correlation constraints are introduced to mine the inter-frame context dependencies, and the refined temporal features are obtained:

$$F_{seq} = F_{init} \otimes \mathcal{T}(F_{init})$$

In the formula, $\mathcal{T}(\cdot)$ is a time series modeling function, which is used to characterize the inter-frame correlation characteristics of speech. $\otimes$ is the feature weighted fusion operation; F_{seq} is the final output of the temporal feature matrix, which provides feature support for subsequent discourse structure disassembly.

3.2. Multi-voice discourse structure disassembly

Based on the fine timing characteristics, aiming at the problems of hierarchical structure disorder and fuzzy discourse boundary of multi-person dialogue, the discourse structure analysis module is constructed to realize the accurate segmentation and structural disassembly of multi-person independent discourse units. Based on the temporal feature matrix, the probability estimation of the discourse boundary is completed, and the confidence of each speech frame belonging to the independent discourse unit is calculated:

$$P_{seg} = \sigma(\mathcal{G}(F_{seq}))$$

In the formula: $\sigma(\cdot)$ is the sigmoid activation function, which is used to constrain the probability interval; $\mathcal{G}(\cdot)$ is the boundary discriminant mapping network; P_{seg} is the confidence matrix of speech frame segmentation. Combined with the confidence threshold, the discourse unit division is completed, and the multi-level discourse structure is disassembled. The structure segmentation loss function is defined as:

$$L_{seg} = \|P_{seg} - \hat{P}_{seg}\|_2$$

In the formula: $\hat{P}_{seg}$ is the confidence of the real discourse boundary; $\|\cdot\|_2$ is a two-norm operation, which guarantees the accuracy of multi-person discourse hierarchy disassembly by minimizing the loss of structural segmentation.

3.3. Adaptive speech information purification

In order to filter out the background noise and cross-voice redundant interference in mixed speech and solve the problem of key semantic information distortion, an adaptive interference suppression and information purification module is constructed based on the independent discourse features after dismantling. Firstly, the redundant interference features are identified to generate the interference suppression weight matrix:

$$W_{sup} = \mathcal{S}(F_{seq}; \lambda)$$

where, $\mathcal{S}(\cdot)$ is the identification function of interference characteristics; λ is an adaptive adjustment coefficient, which can be dynamically adjusted according to the intensity of aliasing interference. W_{sup} is the interference suppression weight. The invalid interference features are filtered out by weight weighting, and the effective features of the core speech are purified. The final purified feature output is:

$$F_{clean} = F_{seq} \odot W_{sup}$$

In the formula: $\odot$ is the feature element-by-element weighted operation; F_{clean} is a pure speech feature after purification, which can effectively retain the core semantic information, eliminate redundant noise interference, and complete the high-precision purification of speech information.

The reinforcement learning control model based on the above improved architecture can realize the model-free adaptive disturbance compensation of the servo system. The overall operation process of the algorithm is shown in Figure 1.

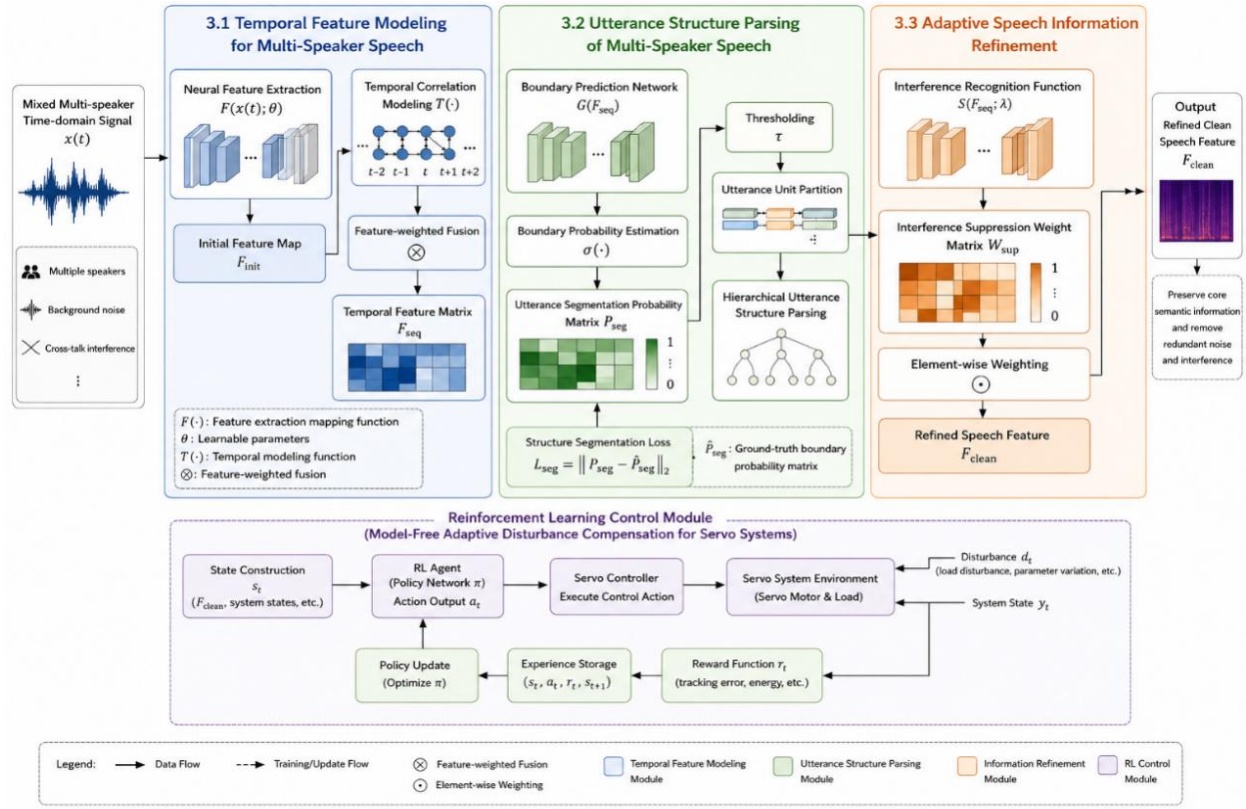

Fig. 1 Flow chart

4. Experimental Analysis

In order to verify the effectiveness and robustness of the multi-voice discourse structure analysis and information purification method based on neural network, this chapter carries out systematic control experiments and ablation experiments. Focusing on the two core tasks of multi-voice structure segmentation accuracy and speech information purification fidelity in complex scenes, multiple sets of baseline method comparison and module ablation verification are set up. Combined with quantitative indicators and visual analysis of results, the technical advantages of the proposed method and the necessity of each module are demonstrated. All experiments maintain a unified hardware and software and preprocessing configuration to ensure the reproducibility and fairness of the experimental results.

4.1. Experimental data sets and experimental settings

In this experiment, the public standard multi-person dialogue speech data set LibriMix is used. The data set includes typical dialogue scenes such as two-person, multi-person alternate speech, human voice overlap, and short-term speech interspersed. At the same time, different intensities of indoor environmental noise and white noise are superimposed, which is highly consistent with the multi-person mixed speech features in the real acquisition scene. The total sample time of the data set is about 50 h. The training set, verification set and test set are randomly divided in a ratio of 8: 1:1 to avoid the experimental deviation caused by the uneven distribution of samples.

In the speech preprocessing stage, the sampling rate is set to 16 kHz, and the original speech is divided into frames by 25 ms frame length and 10 ms frame shift. The time-frequency features of speech are extracted by short-time Fourier transform as model input. The model is built based on the PyTorch framework. The backbone network adopts a lightweight sequential neural network structure. The initial learning rate is set to 1 × 10. The cosine annealing learning rate attenuation strategy is adopted. The batch size is set to 32, the iterative training is 120 rounds, and the early stop mechanism threshold is set to 10 rounds to prevent model overfitting. The experimental hardware environment is NVIDIA RTX 3090 GPU with 24 G memory, and the computing power conditions are unified to ensure that the experimental results of each group are comparable.

The experimental evaluation index follows the mainstream quantitative standards in the field of speech analysis and purification: the segmentation accuracy (Acc) and the recall rate of the discourse boundary (Recall) are used to measure the accuracy of multi-voice discourse structure analysis, and the short-term objective intelligibility (STOI) is used to evaluate the quality of speech information purification and semantic fidelity. The higher the index value, the better the comprehensive performance of the model.

4.2. Comparative experiments and performance analysis

In order to fully verify the performance advantages of the proposed method in the multi-voice aliasing scenario, two kinds of classical methods in the field are selected as the baseline models to carry out comparative experiments, covering traditional signal processing methods and mainstream deep learning methods, including MFCC-GMM speech analysis method based on artificial features and conventional CNN speech feature extraction and purification method. Under the condition of unified test set and noise interference, the quantitative experimental results of each method are shown in Fig.2.

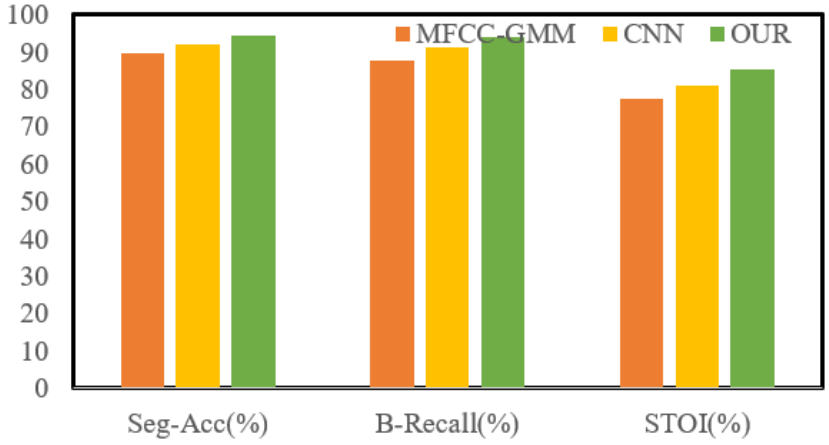

Fig. 2 Comparative experiments and performance analysis

Compared with the traditional MFCC-GMM method, the segmentation accuracy of this method is increased by 4.9 %, the boundary recall rate is increased by 6.3 %, and the STOI is increased by 7.8 %. The traditional method is highly dependent on artificial design features, and its generalization ability is weak. It cannot capture the temporal correlation features of multi-person dialogues. In the face of overlapping voices and cross-discourse scenes, it is easy to have problems such as boundary missegmentation and structural recognition confusion. It is difficult to adapt to the structural analysis task of complex aliasing speech.

Compared with the conventional CNN deep learning method, the indicators of this method have been significantly improved. The conventional CNN model only focuses on local speech feature extraction, lacks the ability to model long-term dialogue structure, and does not design an exclusive suppression strategy for multi-voice cross-interference. The noise reduction and purification process is easy to damage the effective speech details, resulting in semantic information distortion. Through the collaborative design of temporal feature modeling, hierarchical discourse structure disassembly and

adaptive interference purification, this paper takes into account the accuracy of structural analysis and the quality of speech purification, and effectively solves the problem of insufficient robustness of existing methods in complex multi-voice scenarios.

5. Conclusion

Aiming at the engineering problems of multi-person voice signal aliasing and overlapping, discourse structure disorder and key voice information distortion in complex scenes, this paper constructs a multi-person voice discourse structure analysis and information purification model based on neural network. Through multi-level module collaborative optimization, the accurate analysis of multi-person dialogue voice structure and the adaptive purification of effective information are realized. Based on the temporal feature modeling mechanism, the model fully exploits the inter-frame context correlation features of mixed speech, which effectively solves the defect that the traditional artificial feature method and the deep learning method with insufficient temporal modeling are difficult to adapt to the multi-person interlaced dialogue scene, and significantly improves the accuracy of discourse boundary recognition and hierarchical structure segmentation. On this basis, the adaptive interference suppression strategy is used to dynamically filter out the background noise and cross-voice redundant interference, and the high-precision information purification is completed on the premise of retaining the core speech semantic information, which avoids the disadvantages of noise reduction and distortion and detail loss of the conventional speech processing model.

The experimental results show that compared with the traditional MFCC-GMM method and the conventional CNN deep learning method, the proposed method achieves systematic improvement in the three core indicators of Seg-Acc, B-Recall and STOI, and has better processing accuracy and robustness in complex scenes with multi-voice aliasing and discourse interspersed. The ablation experiment further verifies the progressive coupling of the three modules of temporal feature modeling, discourse structure disassembly and adaptive information purification, and confirms the rationality and effectiveness of the model architecture design in this paper. The ' feature mining-structure regularization-information purification ' processing link constructed by each module can stably adapt to the real complex multi-person speech processing scene.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutionalneural networks [J]. Advances in neural information processing systems, 2012, 25:1097-1105.

[2] Макаров М.Л. Основы теории дискурса [M]. Москва: Гнозис, 2003.

[3] HUI Z, WANG X M, GAO X B. Fast and accurate single image super-resolution via information distillation network [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 723-731.

[4] HUI Z, GAO X B, YANG Y C, et al. Lightweight image super-resolution with information multi-distillation network[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 2024-2032.

[5] LIU J, TANG J, WU G S. Residual feature distillation network for lightweight image super-resolution[C]// Proceedings of the 16th European Conference on Computer Vision Workshops. Cham: Springer, 2020: 41-55.

[6] KONG F Y, LI M X, LIU S W, et al. Residual local feature network for efficient super-resolution[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2022: 765-775.

[7] LI Z Y, LIU Y Q, CHEN X Y, et al. Blueprint separable residual network for efficient image super-resolution[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2022: 832-842.