mathematics

MDPI

Article

Improving Speaker Diarization for Overlapped Speech with Texture-Aware Feature Fusion

Chengli Sun [1,2], Miao Sun [1,*] and Wenrui Wei [2]

[1] School of Artificial Intelligence, Guangzhou Maritime University, Guangzhou 510725, China; sunchengli@gzmtu.edu.cn
[2] School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China; wenrui_wei@163.com
* Correspondence: sunmiao@gzmtu.edu.cn

Abstract

Speaker diarization (SD), which aims to address the "who spoke when" problem, is a key technology in speech processing. Although end-to-end neural speaker diarization methods have simplified the traditional multi-stage pipeline, their capability to extract discriminative speaker-specific features remains constrained, particularly in overlapping speech segments. To address this limitation, we propose EEND-ECB-CGA, an enhanced neural network built upon the EEND-VC framework. Our approach introduces a texture-aware fusion module that integrates an Edge-oriented Convolution Block (ECB) with Content-Guided Attention (CGA). The ECB extracts complementary texture and edge features from spectrograms, capturing speaker-specific structural patterns that are often overlooked by energy-based features, thereby improving the detection of speaker change points. The CGA module then dynamically weights the texture-enhanced features based on their importance, emphasizing speaker-dominant regions while suppressing noise and overlap interference. Evaluations on the LibriSpeech_mini and LibriSpeech datasets demonstrate that our EEND-ECB-CGA method significantly reduces the diarization error rate (DER) compared to the baseline. Furthermore, it outperforms several mainstream end-to-end clustering-based approaches. These results validate the robustness of our method in complex, multi-speaker environments, particularly in challenging scenarios with overlapping speech.

Keywords: speaker diarization; end-to-end learning; texture features; feature fusion; attention mechanism

MSC: 68T07; 68T10

Academic Editors: Shixi Hou and Yundi Chu

Received: 13 November 2025
Revised: 4 December 2025
Accepted: 8 December 2025
Published: 11 December 2025

Citation: Sun, C.; Sun, M.; Wei, W. Improving Speaker Diarization for Overlapped Speech with Texture-Aware Feature Fusion. *Mathematics* **2025**, *13*, 3950. https://doi.org/10.3390/math13243950

1. Introduction

Advances in speech recognition have enabled the proliferation of intelligent devices, such as smart speakers and mobile phones, which leverage voice interfaces to facilitate human–computer interaction. In real-world scenarios, audio streams often contain speech from multiple speakers. Direct recognition of such audio may include non-target speech, which detrimentally impacts recognition accuracy. Speaker Diarization (SD), also referred to as speaker segmentation or turn-taking detection, can effectively address this by partitioning an input audio stream into homogeneous segments according to speaker identity [1]. This technology has demonstrated high application value in fields such as video conferencing, in-vehicle voice interaction, and multimedia content analysis [2,3].

Traditional speaker diarization methods primarily follow a multi-stage processing pipeline comprising voice activity detection (VAD), speaker segmentation, speaker embedding extraction, and clustering. Within this paradigm, speaker representations have evolved from MFCCs to i-vectors and x-vectors [4–6], while clustering techniques have employed K-means, Agglomerative Hierarchical Clustering (AHC), and spectral clustering [7–10]. To refine performance, statistical models like the Hidden Markov Model (HMM) [11] were introduced, with the VBx method [12] demonstrating notable effectiveness by applying a variational Bayesian HMM to x-vector sequences. Other significant contributions include the GMM-UBM model [13] and the streaming SF-EV model [14]. Despite extensive research, these cascaded frameworks still suffer from complex training procedures and cumulative error propagation across modules. More critically, their inherent design struggles to handle overlapping speech effectively, leading to performance bottlenecks in realistic conversational settings [15].

To overcome these limitations, deep learning-based end-to-end neural diarization (EEND) approaches have emerged. Early deep learning models for speaker diarization utilized Bidirectional Long Short-Term Memory (BLSTM) networks [16]. With the introduction of the attention mechanism, self-attention-based EEND models became the mainstream approach [17,18]. However, the direct application of these early end-to-end models was prone to overfitting and remained insufficiently effective in handling speaker overlap. To address this more effectively, Horiguchi et al. proposed an extension of EEND based on Encoder–Decoder Attractors (EEND-EDA) [19]. This method first employed a Transformer encoder and used an LSTM-based encoder–decoder to generate multiple attractors. Each attractor is then multiplied by the embeddings generated by EEND to calculate the speech activity for different speakers. Wang et al. improved the loss function of EEND-EDA, proposing an end-to-end method with Absolute Speaker Loss (LS-EEND) [20]. Additionally, Kinoshita et al. proposed the end-to-end neural diarization with vector clustering (EEND-VC) method [21], which integrates end-to-end modeling with clustering-based diarization principles to combine the strengths of both approaches. Härkönen et al. proposed the EEND-M2F method that employs the masked attention mechanism to generate speaker activity masks directly [22].

Building upon EEND-VC, researchers have attempted to integrate various clustering algorithms. For instance, Delcroix et al. combined the variational Bayesian HMM clustering algorithm for x-vectors (VBx) [12] with the EEND model, proposing the Multi-Stream VBx (MS-VBx) model [23]. While such approaches strengthen the role of clustering and reduce output variability, their performance remains sensitive to dataset scale and suffers in the presence of multiple speakers and overlaps due to simplified clustering mechanisms. Parallel efforts have sought deeper integration with deep neural networks (DNNs), yielding methods such as the Unbounded Interleaved-State RNN (UIS-RNN) [24], SSGD [25], and Target Speaker VAD (TS-VAD) [26]-the latter two having demonstrated notable effectiveness. More recently, Ahmed et al. introduced the Neuro-TM Diarizer [27] that integrates MarbleNet, TitaNet, and a time-delay neural network that further enhanced the performance of speaker diarization. The powerful deep learning principles driving these EEND advancements have also demonstrated significant potential across diverse domains such as neutron-gamma discrimination [28] and one-shot talking head generation [29].

Despite these significant advances, end-to-end methods face several persistent challenges. These include limited generalizability to speaker counts beyond the training distribution, high computational costs from iterative inference or global attention, significant latency from sequential decoders, and a heavy reliance on large-scale simulated data that often leads to performance degradation in cross-domain scenarios. Crucially, a common shortcoming across many existing approaches is their primary focus on mod-

eling global temporal context, while neglecting the local acoustic structures and texture details embedded in the speech spectrogram. These fine-grained features are essential for characterizing speaker individuality. This insufficient mining of discriminative local information ultimately limits the model's capability in highly overlapping and complex acoustic environments.

To address this gap, we propose an end-to-end speaker diarization method enhanced by a texture feature fusion module. The novel contributions are summarized as follows:

- Design of the ECB-CGA module: A texture-aware feature fusion module is proposed, adopting a multi-branch convolutional design inspired by image processing techniques. It leverages Sobel and Laplacian operators to extract complementary texture and edge features, enhancing the model's ability to capture discriminative patterns in overlapping speech.
- Dynamic attention mechanism: A dynamic attention mechanism is introduced, which adaptively allocates weights based on input characteristics. This improves model robustness in overlapping speech scenarios by prioritizing critical temporal-spectral regions.
- Integration into EEND-VC framework: The ECB-CGA module is integrated into the EEND-VC framework to construct the EEND-ECB-CGA model. This integration effectively improves diarization performance, particularly in complex acoustic environments with overlapping speech.

The rest of the paper is organized as follows: Section 2 (Methods) reviews the EEND-VC framework and details the proposed method. Section 3 (Experimental setup) outlines the experimental configuration and datasets. Section 4 (Results and Discussion) presents and discusses the empirical findings, while Section 5 (Conclusions) summarizes the study.

2. Methods

While most existing speaker diarization methods focus on the time–frequency characteristics of speech signals, they often overlook inherent structural information embedded in the spectrogram. To address this problem, we enhance the end-to-end model within the End-to-End Neural Diarization with Vector Clustering (EEND-VC) framework by integrating an Edge-Oriented Convolution Block (ECB) into the Content-Guided Attention (CGA) mechanism, forming a novel ECB-CGA feature fusion module. This integration enables the model to capture richer structural and multi-scale representations, thereby improving diarization accuracy. We begin by reviewing the baseline EEND-VC method and subsequently detail our proposed approach.

2.1. EEND-VC Method

The EEND-VC framework consists of two parts: neural network training and clustering computation, as illustrated in Figure 1. To manage computational complexity, the input audio is first divided into fixed-duration, non-overlapping chunks. For each chunk, frame-level feature sequences are extracted and processed by a neural network to estimate speech activities. This separation task is formulated as a multi-label classification problem and handled through a series of Transformer encoders and linear layers. After the separation task, speaker embeddings are further computed and extracted.

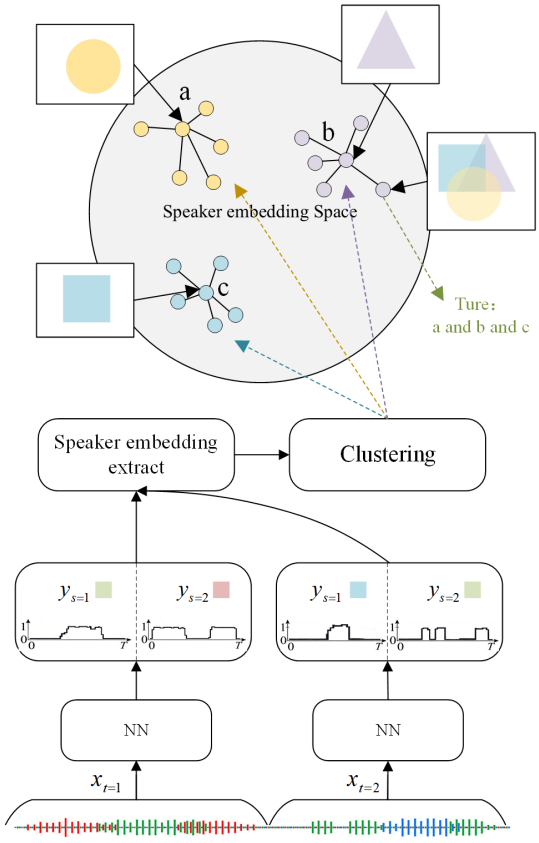

Figure 1. Block diagram of the EEND-VC system.

For the i-th chunk, the network first extracts frame-level features $h_{t,j}$ through an encoder, then outputs the activity probability for the s-th speaker via a linear layer Linear_s^S [21]:

$$\begin{aligned} \hat{y}_{t,i,s} &= \text{sigmoid}\left(\text{Linear}_s^{\text{D}}(h_{t,i})\right) \\ \text{Linear}_s^{\text{D}}(h_{t,i}) &= W_{t,i}h_{t,i} + b_{t,i}. \end{aligned} \tag{1}$$

where $h_{t,i}$ is a D-dimensional internal representation, $\text{Linear}_s^{\text{D}}(\cdot) : \mathbb{R}^{\text{D}} \to \mathbb{R}^1$ is a fully connected layer to estimate the diarization result, and $\text{sigmoid}(\cdot)$ is the element-wise sigmoid function [21].

Simultaneously, the network generates frame-level embeddings $z_{t,i,s}$ through another linear layer Linear_s^S. A chunk-level speaker embedding is obtained by performing a weighted summation using the activity probability [21]:

$$\hat{e}_{t,i,s} = \frac{\sum_{t=1}^{T} \hat{y}_{t,i,s} z_{t,i,s}}{\left\| \sum_{t=1}^{T} \hat{y}_{t,i,s} z_{t,i,s} \right\|} \tag{2}$$

This embedding is normalized to ensure compact distances within the same speaker class and separation between different speakers in the embedding space. EEND-VC employs a multi-task loss function to jointly optimize speaker activity prediction and embedding quality [21]:

$$L = (1 - \alpha) L_{diarization} + \alpha L_{speaker} \tag{3}$$

Here, α is a hyper-parameter to weight the two loss functions that was set at 0.01 according to the parameter setting in [21], $L_{diarization}$ uses permutation-invariant training (PIT) to minimize binary cross-entropy, and $L_{speaker}$ employs contrastive loss. The extracted speaker embeddings are then clustered to associate speaker labels across segments.

Typically, each chunk is assumed to contain speech from two speakers. However, since the total number of speakers in a recording may exceed the number per chunk, and the system can output speakers in any order, speaker ambiguity arises in the EEND stage output. EEND-VC uses the Agglomerative Hierarchical Clustering (AHC) algorithm to cluster speaker embeddings to resolve this ambiguity, associating labels across chunks and estimating the correct number of speakers. This method reduces memory consumption through chunking and utilizes embedding clustering to achieve long-term contextual consistency, providing a solution that balances efficiency and accuracy for complex conversational scenarios.

A key limitation of this approach, however, is its heavy reliance on the accuracy of the speaker embeddings extracted by the EEND model. This reliance poses a problem because conventional clustering methods struggle with speech segments containing overlap, leading to a significant degradation in performance. As shown in Figure 1, 'NN' denotes the aforementioned neural network process, where, during clustering, three different overlapping features are incorrectly grouped into one class. To mitigate this, we introduce a novel Edge-Oriented Convolution Block and Content-Guided Attention (ECB-CGA) module to enhance speaker feature representation.

2.2. Feature Fusion Module

2.2.1. Content-Guided Attention

The Content-Guided Attention (CGA) module [30] is an advanced attention mechanism in deep learning designed to enhance the network's focus on task-relevant regions, thereby improving feature representation. The detailed process of CGA is illustrated in Figure 2. Initially applied in image dehazing, it performs weighted processing on input feature maps. The module consists of three core components: spatial attention, channel attention, and weight refinement, generating channel-specific Spatial Importance Maps (SIMs) in a coarse-to-fine manner. Let $X \in \mathbb{R}^{C\times H\times W}$ represent the input feature map. The goal of CGA is to generate channel-specific SIMs ($W \in \mathbb{R}^{C\times H\times W}$) with the same dimensions as X.

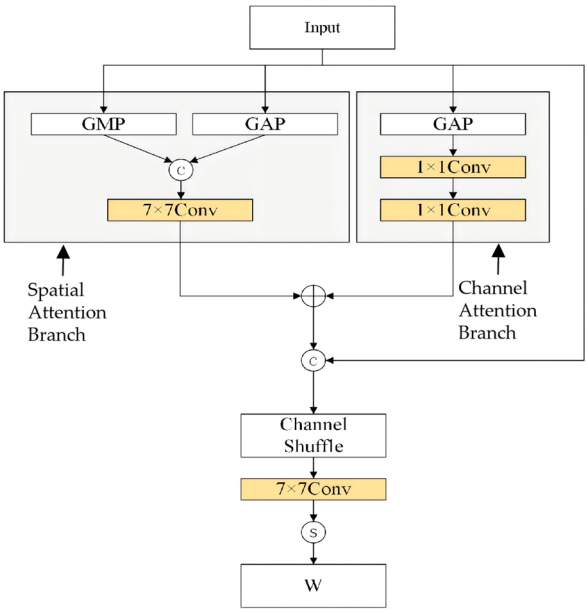

Figure 2. Structure diagram of the CGA module.

2.2.2. Spatial Attention

In the spatial attention branch, Global Max Pooling (GMP) and Global Average Pooling (GAP) are first applied to the input feature X. GMP selects the maximum value along

the spatial dimensions for each channel, highlighting the most significant activation locations, while GAP averages each channel over the spatial dimensions, providing smoother global statistics. The outputs are concatenated along the channel dimension to form a complementary representation of spatial information, providing richer input for subsequent convolution operations. After pooling and concatenation, a 7×7 convolution layer is used to transform the concatenated features. The output of this convolution can be regarded as a preliminary estimate of the spatial distribution, denoted as W_s, which will be fused with the output from the channel attention branch later.

2.2.3. Channel Attention

The structure of the channel attention module includes an average pooling layer, two 1×1 convolution layers, and a ReLU activation function. The channel attention branch first performs global average pooling on the input feature X along the channel dimension. Specifically, this operation averages the data of each channel over the spatial dimensions, resulting in a one-dimensional vector equal in length to the number of channels, representing the overall activation level of each channel. After obtaining the channel-wise average pooling result, a nonlinear mapping is performed using a 1×1 convolution and ReLU activation function. Subsequently, another 1×1 convolution layer maps the channel number back to the original dimension, yielding the channel attention map W_c. This map reflects the importance of each channel for the overall feature representation.

2.2.4. Weight Refinement

After completing the spatial and channel attention branches, the CGA module fuses the generated attention maps W_s and W_c through addition, obtaining a preliminary fused attention map W_{cos}. This map is then concatenated with the original features or certain intermediate features along the channel dimension to combine information from different dimensions. Next, a channel shuffle operation is applied to the concatenated result to encourage cross-channel information exchange, followed by a 7×7 Group convolution layer. Group convolution reduces the number of parameters while enhancing the processing capability for local channel groups, allowing more targeted extraction of features within the same group. This further refines the attention information and improves the network's learning of inter-channel dependencies. The output of the group convolution is normalized by a sigmoid activation function, producing the final attention weight W. Each value of this weight falls within the range $[0, 1]$, indicating the importance of that location for subsequent feature learning. By combining with the input feature through element-wise multiplication or other methods, the network can adaptively focus on more critical regions or channels, achieving refined feature recalibration.

2.2.5. Feature Fusion Refinement

The CGA module processes high-level features from the spatial branch and low-level features from the channel branch, which are denoted as F_{high} and F_{low}, respectively. The weight for the low-dimensional feature is denoted as W, and for the high-dimensional feature, it is denoted $1 - W$. The product of the weights and features of different dimensions is summed with the original feature, and then a 1×1 convolution $C_{1 \times 1}$ is used to map the features back to the original dimension, obtaining the final fused feature. Feature refinement and fusion are performed as follows [30]:

$$F_{\text{fuse}} = C_{1 \times 1}\left(F_{\text{low}} \cdot W + F_{\text{high}} \cdot (1 - W) + F_{\text{low}} + F_{\text{high}}\right) \tag{4}$$

This method preserves original information to a great extent while acquiring features from more dimensions, which is beneficial for subsequent embedding feature extraction.

2.3. Edge-Oriented Convolution Block (ECB)

The ECB module [31], as shown in Figure 3, is an efficient multi-branch parallel convolution structure designed to extract rich local information and edge gradients from input features. The ECB module employs a multi-branch parallel architecture inspired by multi-scale feature extraction techniques used in signal processing and computer vision. It utilizes standard convolutions to extract basic textures, complemented by Sobel and Laplacian operators that specifically capture edge and second-order structural features, respectively. The outputs of each branch are fused to generate highly discriminative features that preserve global semantics while enhancing local details. The ECB module contains five parallel branches, denoted as T_1 to T_5, each performing different convolution or filtering operations. The output y of the ECB layer can be expressed by the following formula:

$$y = y_1 + y_2 + y_3 + y_4 + y_5 \tag{5}$$

where y_1, y_2, y_3, y_4, and y_5 represent the outputs of the five branches, respectively.

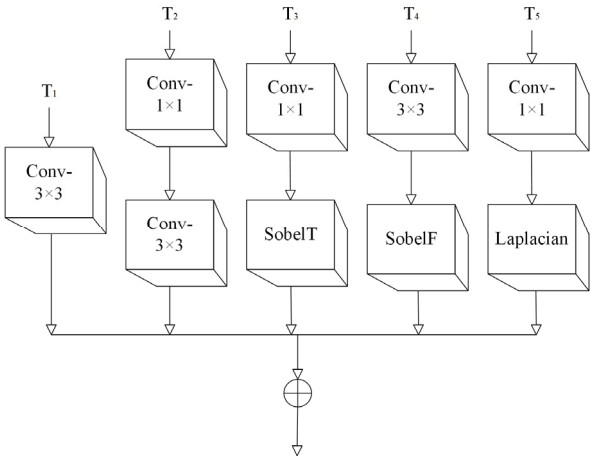

Figure 3. Structure diagram of the ECB module.

The five parallel branches constitute a multi-resolution feature extraction system. T_1 is a 2D convolution with a 3×3 kernel and padding of 1, which can extract basic texture features within a local range, preserving spatial resolution while performing smooth aggregation of neighborhoods. Every other branch contains two convolution blocks, and the first convolution block in each layer is a 2D convolution with a 7×7 kernel.

The second convolution block in T_2 is a 2D convolution with a 3×3 kernel. Let X represent the input to the ECB layer. The T_2 layer can be expressed as:

$$y_2 = \text{conv3}(\text{conv1}(X)) \tag{6}$$

In Equation (6), $\text{conv3}(\cdot)$ and $\text{conv1}(\cdot)$ denote 3×3 convolution and 1×1 convolution, respectively.

T_3 and T_4 use custom convolution kernels as the second convolution to capture gradient information in different directions. The custom kernels can be represented by:

$$W_{D_T} = S_{D_T} \cdot D_T,\ W_{D_F} = S_{D_F} \cdot D_F \tag{7}$$

where W_{D_T} and W_{D_F} represent the second convolution kernels for T_3 and T_4, respectively. S_{D_T} and S_{D_F} represent the learnable parameters for the second convolution kernels of T_3 and T_4. D_T and D_F represent the first-order edge feature extraction along the time and

frequency dimensions of the speech spectrogram using Sobel operators, identifying local directional changes. D_T and D_F are defined in Equation (8):

$$D_T = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix}, D_F = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{8}$$

Then, the T_3 and T_4 layers can be expressed as:

$$y_3 = W_{D_T} \bigotimes (\text{conv1}(X)) + B_{D_T},\ y_4 = W_{D_F} \bigotimes (\text{conv1}(X)) + B_{D_F} \tag{9}$$

where $\otimes$ denotes the convolution operation. B_{D_T} and B_{D_F} represent the initial biases for the T_3 and T_4 convolution layers, respectively.

The T_5 employs a Laplacian operator D_{lap}, capturing second-order structural variations critical for characterizing complex textures. This layer can further mine detailed structures such as inflection points and corner points, exhibiting good discriminative ability for complex textures. The custom kernel can be represented by:

$$W_{lap} = S_{D_{lap}} \cdot D_{lap} \tag{10}$$

where W_{lap} represents the second convolution kernel for the T_5, $S_{D_{lap}}$ represents the learnable parameter for the second convolution kernel of T_5, and D_{lap} represents the second-order edge feature extraction of the speech spectrogram using the Laplacian operator, implementing second-order gradient detection through a second-order difference form, helping to capture subtle structures like inflection points and corners. Then, the T_5 layer can be expressed as:

$$y_5 = W_{D_{lap}} \bigotimes (\text{conv1}(X)) + B_{D_{lap}} \tag{11}$$

where $B_{D_{lap}}$ represents the initial bias for the T_5 convolution layer and D_{lap} is defined as:

$$D_{lap} = \begin{bmatrix} 0 & +1 & 0 \\ +1 & -4 & +1 \\ 0 & +1 & 0 \end{bmatrix} \tag{12}$$

The multi-operator parallel structure of ECB processes input features through multi-angle filtering in the same layer, enabling the capture of significant edges, corners, and other high-frequency information, as well as appropriate aggregation of low-frequency or smooth regions, thereby generating richer texture feature representations.

2.4. Texture Feature Fusion Module

In the original EEND-VC framework, the channel attention mechanism weighted channels based on global statistics, often lacking sensitivity to subtle differences in local regions of the feature map. In practical applications, as the network's demand for multi-scale textures, edges, and global context increases, relying solely on the single weighting strategy of channel attention is insufficient to handle more complex feature distributions. To address this, we design the ECB-CGA module by replacing the channel attention in CGA with an Edge-Oriented Convolution Block (ECB).

As shown in Figure 4, the module starts with frequency-domain feature input, then extracts features of different dimensions through the spatial attention branch and the ECB module. The shape of the 7×7 convolution in the spatial attention branch is adjusted to 2×2. The output of the ECB, together with the spatial attention branch, generates the attention map W. The high-level features are obtained by summing the outputs of the

spatial attention branch and the ECB module, while the low-level features are obtained from the original frequency-domain features. The high and low-level features are fused with different weights W and $1 - W$ using Equation (4). This adaptive scale weighting emphasizes relevant structural information at different resolutions. Then, the final integration is performed at the 1×1 convolution stage. In this study, the ECB module and the CGA module have consistent input/output dimensions, both being (C, H, W) = (1, 256, 500). Here, W is the number of time frames in a single chunk, which is fixed at 500 frames during training, corresponding to 50 s audio with 10 ms frame shift, and is adjustable during inference. H is the feature channel dimension, which is fixed at 256, consistent with the output dimension of the Transformer encoder. C is the channel number. The shape of input X to the ECB-CGA module is (W, H) = (500, 256) (training phase), where X is the frame-level feature map output by the Transformer encoder.

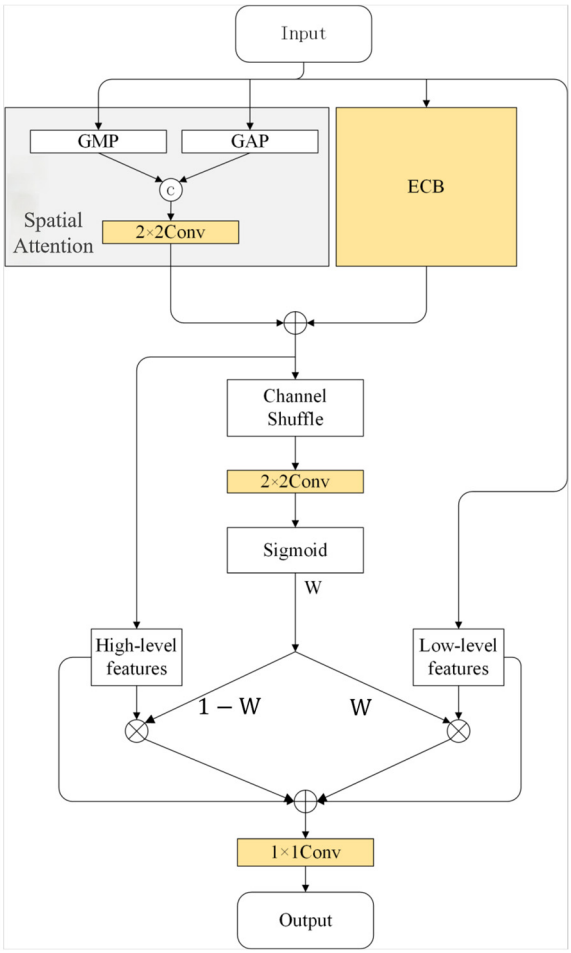

Figure 4. Structure diagram of the ECB-CGA module.

The high-level features typically carry more global semantic information but often lack local precision, while the low-level features contain abundant detailed textures but have insufficient semantic generalization capability. With the new texture feature fusion module, the network is able to retain global semantic information from high-level features while incorporating the detailed local gradient and texture information extracted by the ECB. The low-level features are thus amplified in a targeted manner during fusion. Compared to the original channel attention, the ECB provides richer edge and second-order information, complementing the spatial attention to achieve a superior balance between global and local feature processing.

2.5. End-to-End Speaker Diarization Method Based on Texture Feature Fusion

Based on the idea of texture feature fusion, the ECB-CGA module is integrated into the EEND-VC framework to construct an improved end-to-end speaker diarization (EEND-ECB-CGA) system. The overall structure of the EEND-ECB-CGA model is depicted in Figure 5.

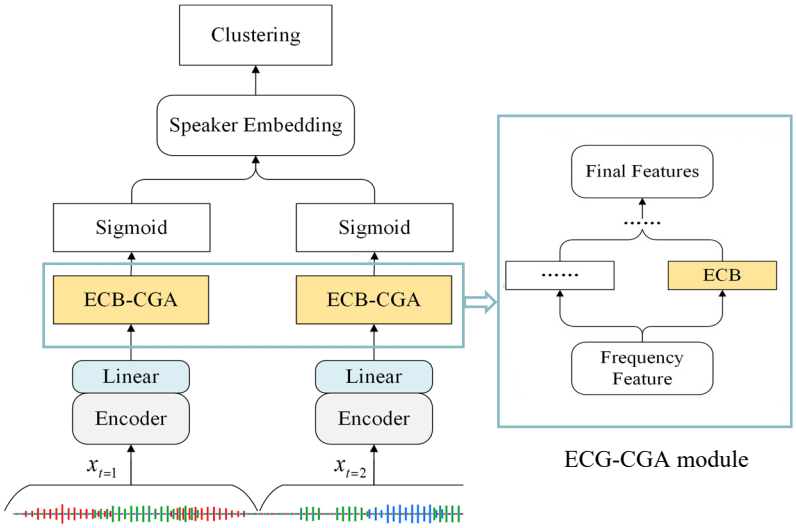

Figure 5. Block diagram of the EEND-ECB-CGA module.

As illustrated in Figure 5, the ECB-CGA module is inserted after the Transformer encoder and a subsequent linear projection layer. This module employs a multi-branch parallel convolution design, integrating conventional convolution, Sobel, and Laplacian operators, among other filtering operations, to extract multi-scale, multi-angle local texture and edge information from the input features. Inside the module, through operations such as convolution, pooling, and channel shuffle, adaptive fusion of the outputs from each branch is achieved, generating a spatial importance map with higher discriminability, providing precise guidance for subsequent feature recalibration. This effectively integrates local details with global semantics, enhancing the robustness and discriminative capability of the overall feature representation.

After completing the speech separation task, the system further computes and extracts speaker embeddings within each speech chunk. Since each chunk is typically preset to contain two speakers, while the actual number of speakers in a recording might be larger, ambiguity exists in the EEND stage output. To solve this problem, the Agglomerative Hierarchical Clustering (AHC) algorithm is used to jointly analyze the speaker embeddings extracted from all chunks, thereby accurately estimating the correct associations of speaker labels across chunks. Finally, based on the clustering results, the separation results from each speech chunk are concatenated into a complete diarization output.

By incorporating the texture feature fusion module, the system can fully capture local details and edge information following the encoder layers, compensating for the limitations of traditional channel attention modules in handling overlapping speech and complex backgrounds. This improvement not only enhances the accuracy of speaker embeddings but also makes subsequent clustering computations more stable, providing a more efficient and robust solution for end-to-end speaker diarization.

3. Experimental Setup

3.1. Hardware and Software Configuration

All experiments were conducted on a Linux workstation equipped with an AMD Ryzen 5 2600X Six-Core Processor (3.60 GHz) (AMD, Santa Clara, CA, USA), 16 GB of RAM, and an NVIDIA GTX 1650 SUPER GPU with 8 GB of VRAM (NVIDIA, Santa Clara, CA, USA). Our models were implemented using the Python 3.11 and PyTorch 2.0 deep learning framework.

3.2. Datasets

To comprehensively evaluate the proposed method, we utilized three publicly available datasets: LibriSpeech, its subset LibriSpeech_mini, and VoxConverse.

- LibriSpeech & LibriSpeech_mini [32]: The LibriSpeech dataset is a large-scale corpus of read English speech derived from audiobooks, containing approximately 460 h of clean speech from 1172 speakers. The LibriSpeech_mini dataset is a standard subset of LibriSpeech. As these datasets contain only single-speaker recordings, we followed common practice to simulate two-speaker conversational mixtures for training and evaluation. This is achieved by overlaying speech segments from different speakers, with varying degrees of overlap, and adding background noise and reverberation. To simulate two-speaker conversational audio, segments were randomly selected from different speakers, with noise and reverberation added using the MUSAN dataset [33]. A two-second silent interval was inserted between speaker segments to mimic real dialog pauses. Reverberation was simulated using 1000 impulse responses, and the signal-to-noise ratio (SNR) was randomly chosen from {5, 10, 15, 20} dB. Training, development, and test sets were generated with overlap ratios ranging from 10% to 90%, each containing 500 simulated two-speaker dialogs. The LibriSpeech dataset and its mini version are used for comparing the proposed method and the baseline method in the experiments.
- VoxConverse [34]: The VoxConverse dataset is a challenging, real-world dataset for speaker diarization, consisting of multi-speaker audio clips extracted from YouTube videos. It is divided into a development (dev) set (216 recordings) and an evaluation (eval) set (232 recordings). The recordings range from 22 to 1200 s in duration and contain between 1 and 21 speakers, presenting a more complex and realistic scenario compared to the simulated LibriSpeech data. This dataset was used to assess the generalization capability of the proposed model and to compare the proposed model to the other mainstream methods.

3.3. Training of the EEND-ECB-CGA Model

The proposed EEND-ECB-CGA model was first trained and validated on the LibriSpeech_mini dataset (8 kHz sampling rate) to conserve computational resources. For feature extraction, 23-dimensional log-Mel filterbank coefficients were computed using a 25 ms frame length and 10 ms frame shift. During training, audio was divided into non-overlapping chunks of 500 frames (50 s). At inference, the model used chunk-level processing combined with the AHC algorithm for diarization. Approximately 25% of test chunks contained fewer than two speakers, while the rest had exactly 2 speakers. The AHC was configured with Euclidean distance as the affinity measure and average linkage method, using a distance threshold θ as the stopping criterion, where clustering terminates when the distance between clusters exceeds this preset value. The threshold θ was determined empirically by selecting the value that yielded the best Diarization Error Rate performance on the development set. Regarding the baseline comparison, we note that the original EEND-VC framework typically assumes two speakers per chunk and can

therefore assign clusters directly without AHC, whereas our method employs AHC to handle variable speaker counts in more realistic scenarios.

The EEND-ECB-CGA encoder incorporates two multi-head attention blocks, each with 256 units divided into four heads. Training used the Adam optimizer with a batch size of 64 for 50 epochs, and the final model was obtained by averaging parameters from the last 10 epochs. The learning rate was set to 0.01, with a maximum of two speakers per chunk and a speaker embedding dimension of 256. The same training procedure was applied to the larger LibriSpeech dataset, which contains 460 h of clean speech from 1172 speakers, yielding 20,000 training mixtures (258 h) and 500 validation mixtures (23 h).

4. Results and Discussion

4.1. Comparison with Baseline Systems

To evaluate the performance of speaker diarization, we employed Diarization Error Rate (DER) as the evaluation metric. This metric aggregates errors from missed speaker speech (Miss), false alarms (FA), and speaker confusion (SC), calculated as the sum of these components divided by the total speech time as follows [35]:

$$\text{DER} = \frac{\text{FA} + \text{Miss} + \text{SC}}{\text{Total Duration of Time}} \tag{13}$$

Here, FA is calculated by measuring the time intervals where the system outputs speaker labels but no speaker is present in the ground truth. Miss is calculated by measuring the time intervals where a speaker is present in the ground truth, but the system does not output a speaker label. SC is calculated by measuring the time intervals where the system outputs a speaker label, but the label does not match the ground truth speaker.

We first evaluated the performance on two-speaker mixed speech segments in a clean environment. Table 1 presents the detailed breakdown of DER components on the LibriSpeech_mini dataset for two overlap ratio ranges. The proposed EEND-ECB-CGA method demonstrates a superior balance across all error types compared to the baseline EEND [16] and EEND-VC [21] models. While the original EEND model exhibits high false alarm rates, and EEND-VC suffers from high miss rates, our approach effectively mitigates these extremes. This is particularly evident as the overlap ratio increases from 20 to 30% to 40–50%, where EEND-ECB-CGA maintains a lower speaker confusion rate, underscoring its enhanced capability to discriminate between speakers in challenging overlapping segments.

Table 1. Experimental results of the EEND-ECB-CGA method on the LibriSpeech_mini dataset compared with baselines.

Method	LibriSpeech Mini					
	Overlap Ratio (%)					
	20–30			40–50		
	Miss	FA	SC	Miss	FA	SC
EEND	6.4	22.8	2.1	5.5	25.3	3.5
EEND-VC	24.0	4.7	8.6	33.2	3.8	8.2
EEND-ECB-CGA	16.2	9.4	5.0	18.3	11.5	6.0

The overall DER comparison on the larger LibriSpeech dataset, as given in Table 2, further corroborates these findings. Our method achieves the lowest DER values under both overlap conditions (4.4% for 20–30% overlap and 5.6% for 40–50% overlap), consistently outperforming not only the standard EEND but also the improved EEND-EDA [19] and the strong EEND-VC baseline. This consistent performance gain across datasets of different

scales highlights the robustness of our texture feature fusion approach. The performance improvement of EEND-ECB-CGA can be attributed to its ability to model the multi-scale structural regularity of speaker characteristics.

Table 2. Experimental results of the EEND-ECB-CGA method on the LibriSpeech_dataset compared with baselines.

Method	LibriSpeech	
	Overlap Ratio (%)	
	20–30 DER	40–50 DER
EEND	9.8	9.6
EEND-EDA	7.6	9.6
EEND-VC	4.6	6.0
EEND-ECB-CGA	4.4	5.6

We conducted multiple runs with different random seeds to assess the statistical significance of our results. On the LibriSpeech dataset (20–30% overlap), the DER for EEND-VC was 4.60% ± 0.08%, while our EEND-ECB-CGA achieved 4.41% ± 0.07%. This improvement is statistically significant ($p < 0.05$, paired t-test).

A qualitative analysis of the spectrograms provides intuitive insights into the separation quality. Figure 6 displays the spectrograms of noisy, mixed speaker audio, characterized by significant clutter and energy overlap. In contrast, the diarization results produced by our EEND-ECB-CGA model plotted in Figure 7 show clearer segment boundaries and more distinct speaker activities. The enhanced separation clarity visible in the spectrograms aligns with the quantitative reduction in error rates, particularly in missed detections. This visual evidence confirms that the ECB-CGA module enables more effective audio separation, even though some highly overlapping regions remain challenging.

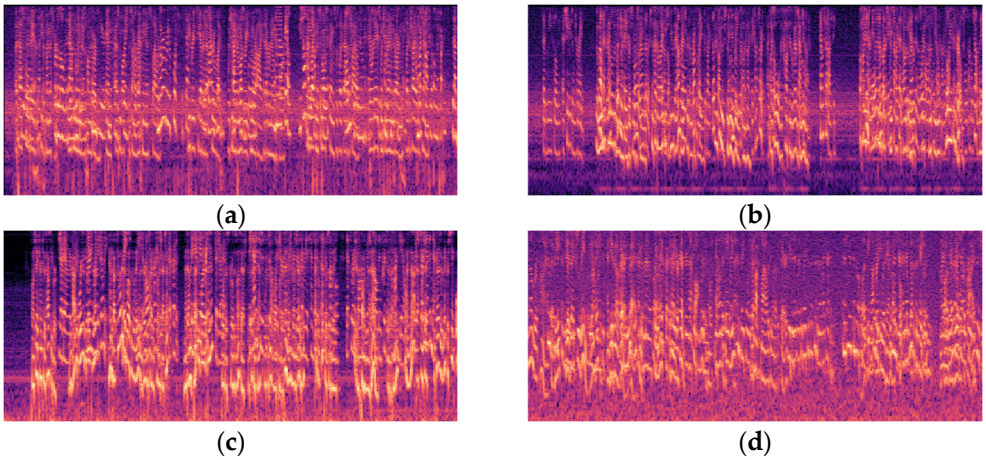

Figure 6. Spectrograms of noisy mixed speaker audio. (**a**–**d**) represent different noisy mixtures, respectively. (**a**) noisy mixture #1; (**b**) noisy mixture #2; (**c**) noisy mixture #3; (**d**) noisy mixture #4.

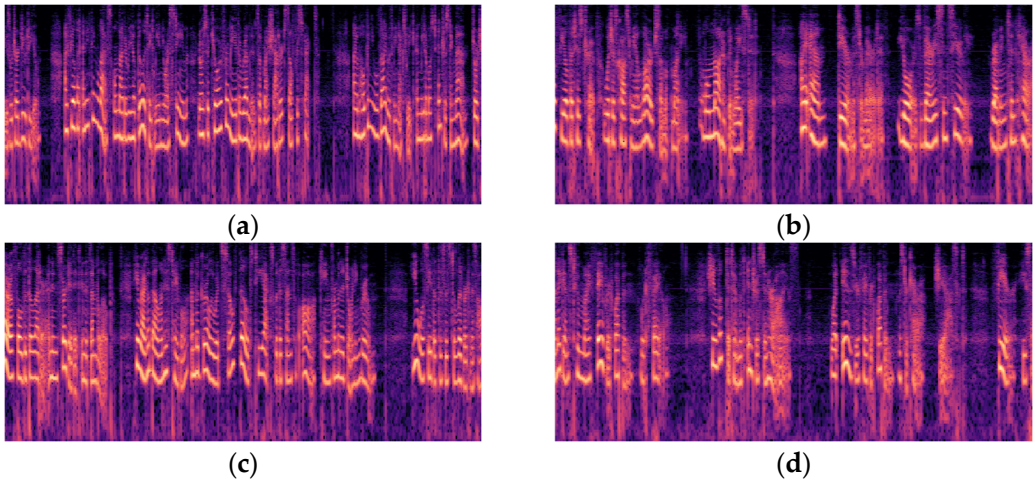

Figure 7. Spectrograms of speaker diarization results for different mixtures. (**a**) diarization results for mixture #1; (**b**) diarization results for mixture #2; (**c**) diarization results for mixture #3; (**d**) diarization results for mixture #4.

4.2. Comparison with Mainstream Methods

To further substantiate the experimental findings, this paper compares the proposed method with other mainstream methods on the VoxConverse dataset, which are considered benchmarks for speaker diarization in overlapping speech scenarios. The mainstream methods include:

- ETDNN + AHC and ETDNN + SC systems: These systems utilize an Extended Time-Delay Neural Network for speaker embedding extraction, followed by either Agglomerative Hierarchical Clustering (AHC) or Spectral Clustering (SC), as configured in [36].
- DiaPer system: A method that integrates visual information with speech features and employs perception guidance to optimize speaker activity detection [37].
- EEND-EDA: An end-to-end model that generates multiple attractors via an encoder–decoder structure [19].
- EEND-M2F: An end-to-end model based on masked-attention transformers that directly generates speaker activity masks without clustering [22].
- The LibriSpeech dataset typically contains only 2–3 speakers, whereas VoxConverse contains more recordings with a higher number of speakers per recording. Comparing the EEND-ECB-CGA system with EEND-M2F, EEND-VC, and other methods, as well as multi-stage methods, it is observed from the results in Table 3 that the EEND-ECB-CGA method reduces the error rate by approximately 2% compared to the baseline (EEND-VC) and shows significant improvements over mainstream methods such as EEND-M2F. The performance gain is primarily due to the ECB-CGA module's dual design principles that address the challenges of overlapping speech scenarios. First, the edge convolution block (ECB) employs Sobel and Laplacian operators to extract fine-grained spectral textures, which are critical for identifying speaker-specific edge patterns in overlapping regions. These operators enable the model to detect subtle acoustic boundaries between speakers that are often obscured in conventional feature extraction methods. Second, the content-guided attention (CGA) mechanism dynamically fuses global contextual information with these local texture features, creating a representation that adapts to complex acoustic environments characterized by a higher number of speakers, overlapping speech, and real-world noise, demonstrating its good generalization capability.

Table 3. Experimental results of the EEND-ECB-CGA system on the VoxConverse_dataset compared with Mainstream Methods.

Dataset	Method	DER
VoxConverse	ETDNN + AHC	13.41
	+ SC	14.02
	DiaPer	22.10
	EEND-EDA	15.75
	EEND-M2F	15.99
	EEND-VC	14.36
	EEND-ECB-CGA	12.64

4.3. Ablation Study

- To validate the effectiveness of the ECB-CGA module, two model variants were designed for ablation experiments:
- EEND-ECB: The ECB module was directly integrated into the Transformer framework of the EEND-VC model, allowing frequency features to be processed directly by the ECB module to capture spectral edge texture features.
- EEND-CGA: The CGA feature fusion module was added to the EEND-VC model, where the untreated frequency features were used as low-dimensional inputs and fused with high-dimensional features.

These two variants, along with the original EEND-VC and the proposed EEND-ECB-CGA, were compared using both the LibriSpeech and LibriSpeech_mini datasets. The results of the ablation study are shown in Figure 8a,b.

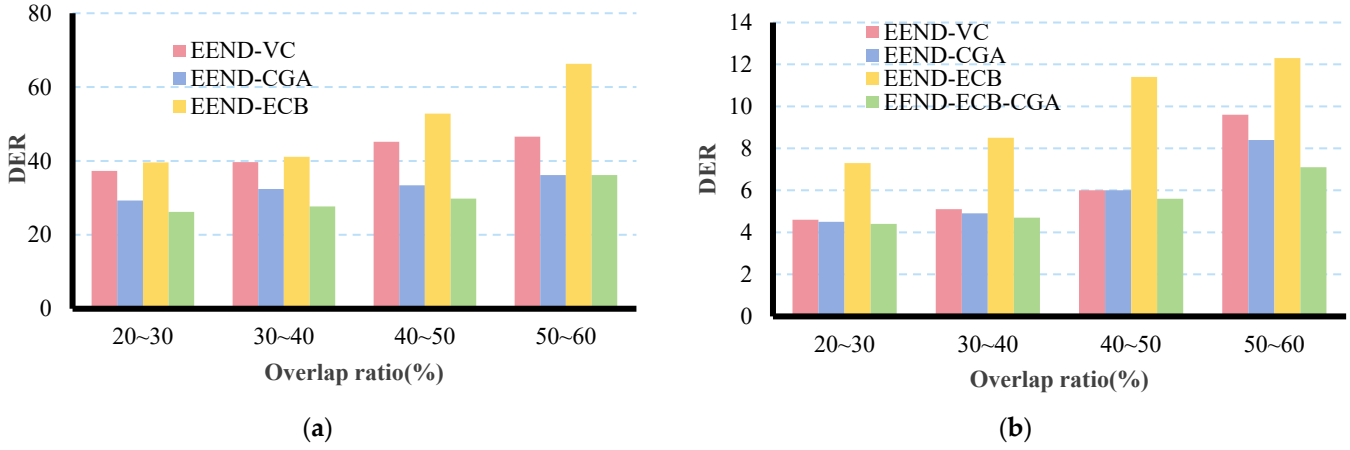

Figure 8. Performance metrics comparison of different models on (**a**) the LibriSpeech_mini dataset and (**b**) the LibriSpeech dataset.

From the figures, it can be observed that compared to the EEND-VC method, EEND-CGA consistently shows a reduction in error rate across all overlap ratios on both datasets, indicating the effectiveness of the CGA feature fusion approach. EEND-ECB, which uses the ECB module alone, exhibits a higher error rate than EEND-VC, particularly in high overlap scenarios (e.g., 50–60% overlap on LibriSpeech_mini, where EEND-ECB's error rate becomes significantly higher than EEND-VC).

It can also be observed from the figures that EEND-ECB-CGA achieves significant improvements over EEND-VC on the LibriSpeech_mini dataset at all overlap ratios. On LibriSpeech, this method also performs better than EEND-VC, with a notable improvement at a 50–60% overlap ratio. These results demonstrate the effectiveness of the CGA approach and the fusion of the ECB-CGA module to enhance speaker diarization performance.

The ablation results reveal a critical synergy between the ECB and CGA modules. The standalone ECB module, specialized in extracting high-frequency edge and texture features, appears to enrich the feature set with fine-grained, locally discriminative information. However, in the absence of the CGA's global contextual refinement, these localized features may lack the structural coherence necessary for stable clustering. Our analysis indicates that the ECB-alone features, while discriminative in a local context, introduce a form of 'high-dimensional noise' or fragmented representations that disrupt the formation of compact, speaker-homogeneous clusters in the embedding space. This aligns with the observed increase in error rate, particularly in high-overlap scenarios where cluster separability is most challenging.

4.4. Computational Complexity Analysis

Due to the relative complexity of the added components in the proposed framework, the proposed EEND-ECB-CGA model increases the per-chunk neural network processing time by approximately 20% compared to EEND-VC. This additional cost is primarily located in the forward pass of the neural network for each audio chunk and can be broken down as follows:

- ECB Module: This module employs five parallel convolutional branches. While the fixed Sobel and Laplacian operators are computationally inexpensive, the multiple standard convolutional layers (especially the 7×7 convolutions) increase the number of floating-point operations (FLOPs).
- CGA Module: This module adds operations for spatial and channel attention, including global pooling, additional convolutions, channel shuffle, and the feature fusion process itself.

Although our method is computationally more expensive than EEND-VC, it achieves a 12.0% relative reduction in DER on the challenging VoxConverse dataset and consistent improvements on LibriSpeech under various overlap conditions. We believe that this substantial boost in accuracy amply justifies the moderate computational increase.

5. Conclusions

This paper has presented EEND-ECB-CGA, an enhanced end-to-end speaker diarization architecture designed to address the persistent challenge of overlapped speech. The proposed method introduces a texture-aware fusion module that integrates Edge-oriented Convolution Blocks (ECB) with Content-Guided Attention (CGA). This design enables the capture of complementary spectro-temporal information, effectively combining fine-grained acoustic texture features and global context features to form more discriminative speaker representations.

Empirical evaluations demonstrate the effectiveness of our approach. The EEND-ECB-CGA model achieved a diarization error rate (DER) of 12.64% on the challenging VoxConverse dataset, representing a significant improvement over the EEND-VC baseline and outperforming other mainstream methods. Consistent performance gains were also observed on the LibriSpeech benchmarks under various overlap conditions. Ablation studies confirmed that the integration of the ECB and CGA modules is crucial for this performance boost, as neither component alone could achieve the same robust results.

This work facilitates robust speaker diarization in real-world applications where overlapping speech and noisy environments are pervasive, thereby addressing a key limitation in prior systems. Future work will explore adaptive clustering mechanisms to better handle recordings with highly variable speaker counts. Additionally, we plan to investigate domain-specific optimizations for particular acoustic environments (e.g., in-car,

far-field) and to refine the model for greater computational efficiency to facilitate its broader industrial deployment.

Author Contributions: Conceptualization, C.S. and M.S.; methodology, C.S. and M.S.; software, C.S.; validation, M.S. and W.W.; writing—original draft, W.W. and M.S.; writing—review and editing, C.S. and M.S.; supervision, C.S.; project administration, C.S.; funding acquisition, C.S. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Key Areas Special Project for Regular Higher Education Institutions in Guangdong Province under Grant 2025ZDZX1023, in part by Ganzhou City Unveiling and Leading Project under Grant 2023ULGX0004, in part by Department of Science and Technology of Jiangxi Province under Grant 20232BCJ22050, Grant CK202404510, and Grant 20252BAC200175, in part by Guangzhou Municipal Education Bureau under Grant 2024312000, Grant 2024312430, in part by Guangzhou Science and Technology Bureau Youth Doctor Sailing Project under Grant 2025A04J4511, in part by Guangdong Key Discipline Research Capacity Building Project under Grant 2024ZDJS056, in part by Natural Science Foundation of Shandong Province under Grant 2025ZDZX1023.

Data Availability Statement: The data presented in this study are available in (LibriSpeech) at (10.1109/ICASSP.2015.7178964), reference number [30], and (VoxConverse) at (10.21437/Interspeech.2020-2337), reference number [31]. These data were derived from the following resources available in the public domain: (https://www.openslr.org/12, accessed on 8 July 2024) and (https://www.robots.ox.ac.uk/~vgg/data/voxconverse/index.html, accessed on 8 July 2024).

Conflicts of Interest: The authors declare no competing interests.

References

1. Tranter, S.E.; Reynolds, D.A. An overview of automatic speaker diarization systems. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1557–1565. [CrossRef]
2. Wang, W.; Qin, X.; Cheng, M.; Zhang, Y.; Wang, K.; Li, M. The DUK-DukeECE diarization system for the voxceleb speaker recognition challenge 2022. *arXiv* **2022**, arXiv:2210.01677.
3. Wang, J.; Du, Z.; Zhang, S. Told: A novel two-stage overlap-aware framework for speaker diarization. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
4. Gelly, G.; Gauvain, J.L. Optimization of RNN-based speech activity detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *26*, 646–656. [CrossRef]
5. Sell, G.; Snyder, D.; McCree, A.; Garcia-Romero, D.; Villalba, J.; Maciejewski, M.; Manohar, V.; Dehak, N.; Povey, D.; Watanabe, S.; et al. Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, 2–6 September 2018; pp. 2808–2812.
6. Sell, G.; Garcia-Romero, D. Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 7–10 December 2014; pp. 413–417.
7. Maciejewski, M.; Snyder, D.; Manohar, V.; Dehak, N.; Khudanpur, S. Characterizing performance of speaker diarization systems on far-field speech using standard methods. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5244–5248.
8. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust DNN embeddings for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.
9. Arora, A.; Raj, D.; Shanmugam Subramanian, A.; Li, K.; Ben-Yair, B.; Maciejewski, M.; Żelasko, P.; García, P.; Watanabe, S.; Khudanpur, S. The JHU multi-microphone multi-speaker ASR system for the CHiME-6 challenge. In Proceedings of the International Workshop on Speech Processing in Everyday Environments (CHiME), Online, 4 May 2020; pp. 48–54.
10. Han, K.; Narayanan, S. A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system. In Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH), Antwerp, Belgium, 27–31 August 2007; pp. 1853–1856.
11. Baum, L.E.; Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **1966**, *37*, 1554–1563. [CrossRef]

12. Landini, F.; Profant, J.; Diez, M.; Burget, L. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Comput. Speech Lang.* **2022**, *71*, 101254. [CrossRef]
13. Reynolds, D.A.; Quatieri, T.F.; Dunn, R. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.* **2000**, *10*, 19–41. [CrossRef]
14. Castaldo, F.; Colibro, D.; Dalmasso, E.; Laface, P.; Vair, C. Stream-based speaker segmentation using speaker factors and eigenvoices. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV, USA, 31 March–4 April 2008; pp. 4133–4136.
15. Efstathiadis, G.; Yadav, V.; Abbas, A. LLM-based speaker diarization correction: A generalizable approach. *Speech Commun.* **2025**, *170*, 103224. [CrossRef]
16. Fujita, Y.; Kanda, N.; Horiguchi, S.; Xue, Y.; Nagamatsu, K.; Watanabe, S. End-to-end neural speaker diarization with self-attention. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Singapore, 14–18 December 2019; pp. 296–303.
17. Fujita, Y.; Kanda, N.; Horiguchi, S.; Xue, Y.; Nagamatsu, K.; Watanabe, S. End-to-end neural speaker diarization with permutation-free objectives. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 15–19 September 2019; pp. 4300–4304.
18. Xue, Y.; Horiguchi, S.; Fujita, Y.; Watanabe, S.; García, P.; Nagamatsu, K. Online end-to-end neural diarization with speaker-tracing buffer. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 841–848.
19. Horiguchi, S.; Fujita, Y.; Watanabe, S.; Xue, Y.; García, P. Encoder-decoder based attractors for end-to-end neural diarization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1493–1507. [CrossRef]
20. Wang, C.; Li, J.; Fang, X.; Kang, J.; Li, Y. End-to-end neural speaker diarization with absolute speaker loss. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dublin, Ireland, 20–24 August 2023; pp. 3577–3581.
21. Kinoshita, K.; Delcroix, M.; Tawara, N. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7198–7202.
22. Härkönen, M.; Broughton, S.J.; Samarakoon, L. EEND-M2F: Masked-attention mask transformers for speaker diarization. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Kos, Greece, 1–5 September 2024; pp. 37–41.
23. Delcroix, M.; Tawara, N.; Diez, M.; Landini, F.; Silnova, A.; Ogawa, A.; Nakatani, T.; Burget, L.; Araki, S. Multi-stream extension of variational Bayesian HMM clustering (MS-VBx) for combined end-to-end and vector clustering-based diarization. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Dublin, Ireland, 20–24 August 2023; pp. 3477–3481.
24. Zhang, A.; Wang, Q.; Zhu, Z.; Paisley, J.; Wang, C. Fully supervised speaker diarization. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6301–6305.
25. Morrone, G.; Cornell, S.; Raj, D.; Serafini, L.; Zovato, E.; Brutti, A. Low-Latency Speech Separation Guided Diarization for Telephone Conversations. In Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2022; pp. 641–646.
26. Medennikov, I.; Korenevsky, M.; Prisyach, T.; Khokhlov, Y.; Korenevskaya, M.; Sorokin, I.; Timofeeva, T.; Mitrofanov, A.; Andrusenko, A.; Podluzhny, I.; et al. Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 274–278.
27. Ahmed, M.; Alharbey, R.; Daud, A.; Ullah Khan, H.; Nawal, J.; Arshad, G. An enhanced deep learning approach for speaker diarization using TitaNet, MarbelNet and time delay network. *Sci. Rep.* **2025**, *15*, 24501. [CrossRef] [PubMed]
28. Zhao, K.; Wang, H.; Wang, X.; An, L.-H.; Chen, L.; Zhang, Y.-P.; Lv, N.; Li, Y.; Ruan, J.-L.; He, S.-H.; et al. Neutron-gamma discrimination method based on voiceprint identification. *Radiat. Meas.* **2025**, *187*, 107483. [CrossRef]
29. Tang, P.; Zhao, H.; Meng, W.; Wang, Y. One-shot motion talking head generation with audio-driven model. *Expert Syst. Appl.* **2025**, *297*, 129344. [CrossRef]
30. Chen, Z.; He, Z.; Lu, Z. DEANet: DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Trans. Image Process.* **2024**, *33*, 1002–1015. [CrossRef]
31. Zhang, X.; Zeng, H.; Zhang, L. Edge-oriented convolution block for real-time super resolution on mobile devices. In Proceedings of the 29th ACM International Conference on Multimedia, Online, 20–24 October 2021; pp. 4034–4043.
32. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.

33. Snyder, D.; Chen, G.; Povey, D. MUSAN: A music, speech, and noise corpus. *arXiv* **2015**, arXiv:1510.08484. [CrossRef]
34. Chung, J.S.; Huh, J.; Nagrani, A.; Afouras, T.; Zisserman, A. Spot the conversation: Speaker diarisation in the wild. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020.
35. Fiscus, J.G.; Radde, N.; Garofolo, J.S.; Le, A.; Ajot, J.; Laprun, C. The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In Proceedings of the International Workshop on Machine Learning for Multimodal Interaction (MLMI), Bethesda, MD, USA, 1–4 May 2006; pp. 369–389.
36. Singh, P.; Ganapathy, S. End-to-end supervised hierarchical graph clustering for speaker diarization. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 448–457. [CrossRef]
37. Landini, F.; Diez, M.; Stafylakis, T.; Burget, L. Diaper: End-to-end neural diarization with perceiver based attractors. *arXiv* **2023**, arXiv:2312.04324. [CrossRef]