Blueprint Separable Subsampling and Aggregate Feature Conformer-Based End-to-End Neural Diarization

Xiaolin Jiao 1, Yaqi Chen 2, Dan Qu 2,* and Xukui Yang 2,*

1 School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China; jxl5056@sina.com
2 School of Information Systems Engineering, Information Engineering University, Zhengzhou 450001, China; chyaqi163@163.com
* Correspondence: qudanqudan@sina.com (D.Q.); gzyangsk@163.com (X.Y.)

Abstract: At present, a prevalent approach to speaker diarization is clustering based on speaker embeddings. However, this method encounters two primary issues. Firstly, it cannot directly minimize the diarization error during the training process; secondly, the majority of clustering-based methods struggle to handle speaker overlap in audio. A viable approach for addressing these issues involves adopting end-to-end speaker diarization (EEND). Nevertheless, training this EEND system generally requires lengthy audio inputs, which must be downsampled to allow efficient model processing. In this study, we develop a novel downsampling layer using blueprint separable convolution (BSConv) instead of depthwise separable convolution (DSC) as the foundational convolutional unit, which effectively preserves information from the original audio. Furthermore, we incorporate multi-scale feature aggregation (MFA) into the encoder structure to combine the features extracted by each conformer block to the output layer, consequently enhancing the expressiveness of the model’s feature extraction. Lastly, we employ the conformer as the backbone network to incorporate the proposed enhancements, resulting in an EEND system named BSAC-EEND. We assess our suggested methodology on both simulated and real datasets. The experiment indicates that our proposed EEND system reduces diarization error rate (DER) by an average of 17.3% for two-speaker datasets and 12.8% for three-speaker datasets compared to the baseline.

Keywords: end-to-end neural diarization (EEND); blueprint separable convolution (BSConv); multi-scale feature aggregation (MFA)

1. Introduction

Speaker diarization is the task of determining “who spoke when” from a given audio input. This technique has various applications, including information retrieval from broadcast news, generation of conference transcripts, and analysis of telephone conversations [1,2]. Moreover, it is a crucial component for implementing automatic speech recognition (ASR) in multi-speaker settings such as telephone conversations [3], conferences [4], and lectures [5]. It has been demonstrated that accurate speaker diarization results can enhance ASR performance by constraining speech masks during the construction of speech-separating beamformers [6,7].

Traditional speaker diarization systems mainly adopt a clustering-based approach [8,9], which sequentially applies the following modules to the input audio to obtain the final results: voice activity detection, voice segmentation, feature extraction, and clustering. However, despite the satisfactory performance of traditional clustering-based systems, they exhibit certain limitations. First, they depend on multiple modules, each requiring individual training. Consequently, clustering-based speaker diarization systems need to be jointly calibrated between different modules, which introduces additional complexity in training. Secondly, although some recent work has attempted to handle situations with multiple speakers at the same time, the clustering-based approach implicitly assumes that each short speech segment has only one speaker speaking, making it difficult to handle overlapping speech [10].
End-to-end speaker diarization presents an optimal solution. Self-attentive end-to-end neural diarization (SA-EEND) [11] is an advanced method designed to model the joint speech activities of multiple speakers. Unlike clustering-based approaches, SA-EEND does not require clustering algorithms; instead, it directly outputs the speaking probabilities for all speakers within each time period. By utilizing a multi-label classification framework, SA-EEND naturally addresses the speaker overlap problem in both training and inference given a multi-speaker audio recording input. Furthermore, SA-EEND follows an end-to-end training strategy, minimizing the diarization error rate and exhibiting remarkable accuracy on a dataset of two-speaker telephone conversations [12].

In this paper, we propose an end-to-end speaker diarization model base on a blueprint separable subsampling layer and an aggregate feature conformer (BSAC-EEND), which enhances system performance by optimizing convolutional subsampling layers and feature encoders. Firstly, the EEND system must process long utterances. During the training phase, to increase efficiency, the system segments input audio, which typically varies in length from a few seconds to several tens of seconds. In contrast, during the testing phase, the entire audio is assessed, leading to longer input data. A critical consideration is the model’s ability to provide robust processing for various utterance durations, particularly for longer inputs. In the realm of speaker recognition, the multi-scale feature fusion conformer [13] has demonstrated proficiency in extracting robust global feature information from speech audios of varying lengths. This is attributed to the model’s ability to capture both global and local features while integrating multi-scale feature representations. Significantly, it exhibits better performance in the case of long-duration test utterances [13]. Consequently, we introduce the MFA structure into the EEND system’s feature extractor. The conformer [14] merges a transformer [15] with convolutional neural networks to effectively capture global and local features, whereas the MFA structure concatenates the output features of all conformer blocks, resulting in multi-scale representations. Secondly, to diminish redundant information in audio frame sequences and enable the model to efficiently process long-duration audio input, downsampling is necessary. Currently, frame stacking subsampling is frequently employed [11,16], although [12] has shown that convolutional subsampling can notably enhance EEND networks’ performance compared to frame-stacked subsampling. Motivated by this finding, we introduce BSConv [17], an advanced version of depthwise separable convolution [18] that more effectively leverages correlations within convolution kernels to achieve efficient convolution separation [17]. We have designed a novel convolution subsampling layer utilizing both unconstrained BSConv (BSConv-U) [17] and subspace BSConv (BSConv-S) [17] structures to optimize the subsampling architecture.

This paper’s primary contributions are threefold: (1) Incorporation of the MFA structure into the EEND model’s feature extractor, thereby enhancing its capability for extracting robust features. (2) Assessment of the performance of the convolutional subsampling layer based on BSConv-U and BSConv-S on both simulated and real data, with results demonstrating superiority over frame stacking subsampling and DSC-based convolutional subsampling. (3) Development of a novel EEND system which employs a conformer as the backbone network, integrates the MFA structure, and incorporates a convolutional subsampling layer based on BSConv. This new EEND system demonstrates significant performance improvements on both simulated and real datasets.

2. Related Work

2.1. Traditional Speaker Diarization Method

Traditional speaker diarization approaches typically comprise a cascaded system consisting of four modules: (1) speech activity detection (SAD), (2) speaker embedding extraction, (3) speaker embedding clustering, and an optional process, (4) overlap processing. Some methods also incorporate an ASR module [19,20]. The majority of research primarily focuses on the second and third modules, namely speaker embedding extraction and speaker embedding clustering. For speaker embeddings, i-vector [21,22], x-vector [23],
and d-vector [24,25] have been investigated. Early works on speaker embedding clustering employed traditional algorithms such as k-means clustering [8,26], aggregated hierarchical clustering (AHC) [9,27,28], mean-shift clustering [21], and spectral clustering [24,29]. More recently, advanced clustering techniques have been proposed, such as Bayesian HMM clustering of x-vector sequences (VBx) [30,31], automatically adjusted spectral clustering [32], and fully supervised clustering [25,33]. These approaches are typically utilized for hard clustering, wherein a sample exclusively belongs to a single class; consequently, most cascaded systems cannot address speaker overlap, with a few exceptions [34]. To handle speaker overlap, some cascaded approaches incorporate the last optional module; however, recent studies occasionally omit this step from the methods and evaluations [22,24,25,32,33]. Additionally, the first module, SAD is often neglected when assessing cascaded methods using annotated speech activities [22,24,25,32,33].

2.2. End-to-End Speaker Diarization

For direct generation of results from audio, one advantage of end-to-end speaker diarization approaches is that they do not need additional modules to detect unvoiced or overlapped voices. Various methods exist for obtaining speaker diarization results within speech separation [35,36]. However, these models are trained on clean speech or time–frequency masks derived from clean speech, making them unsuitable for training on complex audio datasets such as DIHARD [37,38]. In contrast, EEND-based methods, designed to output the posterior probability of multiple speakers speaking simultaneously, can be trained on real data and are better for solving practical problems. The EEND model accepts a sequence of acoustic features, such as MFCC or Fbank, $X = (x_t \in \mathbb{R}^F | t = 1, \ldots, T)$. The neural network produces a speaker label sequence $Y = (y_t | t = 1, \ldots, T)$, wherein $y_t = [y_{t,k} | k \in \{0,1\}, k = 1, \ldots, K]$. It means that at frame speaker is speaking when $y_{t,k} = 1$, and denotes the maximum number of speakers the network can handle. For different speakers $k$ and $k'$, both $y_{t,k}$ and $y_{t,k'}$ may equal one, indicating that speakers $k$ and $k'$ are speaking simultaneously at frame $t$, i.e., overlapping speech. Assuming the output $y_{t,k}$ is conditional independence, the model’s training objective is to maximize the diarization posterior probability $\log P(Y|X) \sim \sum_t \sum_k \log P(y_{t,k}|X)$ of the training data. Multiple candidate reference label sequences $k$ can be generated by exchanging the speaker label $Y$; thus, the loss is computed for all possible reference sequences, and the reference label with the lowest loss is used for error backpropagation. This approach is influenced by the permutation-free objective used in speech separation [39]. Early EEND implementations employed bidirectional long short-term memory (BLSTM) [16], succeeded by self-attention-based EEND networks [11], demonstrating improved DER on two-speaker data, such as the CALLHOME dataset (LDC2001S97) [40] and dialogue audio in the CSJ [41]. Drawing upon the high-performing conformer model in end-to-end continuous speech recognition, literature [12] devised a conformer-based EEND system (CB-EEND), incorporating SpecAugment [42] and a convolutional subsampling layer based on depthwise separable convolution, resulting in significant improvements.

3. BSAC-EEND

This chapter presents the fundamental composition and structure of the proposed BSAC-EEND. As shown in Figure 1, the model’s comprehensive structure encompasses SpecAugment, a BSConv convolution subsampling layer, a conformer block, an MFA structure, a layer norm, a linear layer, and a sigmoid function. The subsequent sections provide detailed descriptions of the model’s critical components.
3.1. Convolution Subsampling Based on BSConv

After in-depth analysis of CNNs, the work [17] came to the conclusion that convolution kernels generally show high redundancy in their depth direction, which means they have a higher intra-kernel correlation. Consequently, the author proposes BSConv, which employs a two-dimensional blueprint to represent each convolution kernel utilizing a weight vector distributed along the depth axis.

Inspired by the correlation within the convolution kernel, DSC assumes that there is high inter-kernel redundancy and adopts a structure which is opposite to BSConv. This, however, has been proven to be less effective when separating convolutions [43].

As a result, we explore the use of BSConv as the fundamental unit of convolutional subsampling and design a novel convolutional subsampling layer to subsample input features. BSConv encompasses two distinct structures: BSConv-U and BSConv-S. Analogous to DSC, they are also composed of pointwise convolution [18] and depthwise convolution [18] with a different order. The depthwise convolution applies a convolution kernel to each channel of the input feature map, resulting in a modification of the output feature map’s shape without altering its channel count. In contrast, the pointwise convolution employs a $1 \times 1$ convolution for every feature map. It exclusively modifies the number of channels, leaving the feature map’s size unaffected. Figure 2 illustrates the transformations in the shape of the input and output. Their structural comparisons to DSC are also displayed in Figure 2.

---

**Figure 1.** BSAC-EEND model.

**Figure 2.** The structure of DSC and BSConv.
Drawing inspiration from these structures, we develop two convolutional subsampling modules, with their respective architectures depicted in Figure 3. We refer to the convolutional subsampling based on BSConv-U as CS-BSCU and the one based on BSConv-S as CS-BSCS. The performance of these BSConv-based convolutional subsampling modules in the EEND system will be discussed in detail in Section 4.

Figure 3. Subsampling layer based on BSConv.

Assuming the input acoustic feature possesses a length of $T$ and a dimension of $F$, denoted as $X \in \mathbb{R}^{T \times F}$, the feature undergoes SpecAugment before being transmitted to the convolutional subsampling layer. This novel convolutional subsampling process, embodied by BSConv, can be described as:

\[
X' = \text{SpecAugment}(X) \tag{1}
\]

\[
X'' = \text{BSConv}(X') \tag{2}
\]

where $X \in \mathbb{R}^{T \times F}$, $X' \in \mathbb{R}^{T' \times F'}$, $T' = T/10$.

3.2. Conformer Block

The self-attention mechanism derived from the transformer effectively captures long-distance global context dependencies; however, it lacks local detail recognition. To model global and local features more directly and efficiently, the conformer model [14] combines both CNN and transformer architectures, thereby better capturing the features necessary for speaker diarization in audio. The principal components of the conformer encoder block encompass the multi-head self-attention (MHSA) and convolution modules. Following the MHSA, the convolution module consists of pointwise convolution, one-dimensional depthwise convolution, and a batch normalization layer succeeding the convolution layer, which aids in training the depth model more readily.
The conformer encoder block’s structure [14] diverges from the transformer [15], incorporating two feedforward neural network (FNN) modules possessing semi-residual connections. These two FNNs are positioned between the MHSA and convolution modules, resembling a macaron structure. Mathematically, for the i-th conformer block and input feature \( E_{i-1} \), the output feature \( E_i \) is calculated as follows:

\[
\hat{E}_i = E_{i-1} + \frac{1}{2} \text{FNN}(E_{i-1}) \\
\hat{E}_i = E_{i-1} + \frac{1}{2} \text{FNN}(E_{i-1}) \\
E''_i = E'_i + \text{Convolution}(E'_i) \\
E_i = \text{LayerNorm}\left(E''_i + \frac{1}{2} \text{FNN}(E''_i)\right)
\]

3.3. MFA Structure

The research in [44,45] demonstrated that low-level features can contribute to high-quality speaker feature extraction. Following this theory, the ECAPA-TDNN [46] system aggregates the output features of all SE-Res2 blocks before the final pooling layer, resulting in a significant performance improvement.

Similarly, in our approach, we concatenate the output features of each conformer block for extracting features from EEND speaker diarization and then pass them to the layer norm layer. Unlike [13], however, we eliminate the pooling layer after the layer norm along with the batch norm layer and other structures. There are two reasons for this: Firstly, the EEND system’s input audio typically contains multiple speakers, making it impossible to obtain representative speaker features using the pooling method. Secondly, the pooling method acquires segment-level representative speaker embeddings, whereas we require frame-level speaker activity analysis, rendering pooling inapplicable. Therefore, we directly use a linear layer and a sigmoid function to output the frame-level speaker activity’s posterior probability. The specific process can be described as:

\[
E = \text{Concat}(E_1, ..., E_i) \\
O = \text{LayerNorm}(E) \\
\text{Output} = \sigma(\text{Linear}(O))
\]

where \( \sigma \) denotes the sigmoid function, \( E_1, ..., E_i \in \mathbb{R}^{T' \times d}, T' = T/10, d \) symbolizes the output dimension of the transformer or conformer encoder, \( E, O \in \mathbb{R}^{T' \times D} \), \( i \) represents the number of encoder blocks, and \( D = d \times i \).

4. Data and Experiments

4.1. Data Preparation

To train the EEND system, we employed the simulated data generation algorithm suggested in [16], which has been used in several studies [11,12,16,47,48]. Firstly, we selected \( N \) speakers, and for each speaker, we chose random speech segments. We then inserted silent parts between these segments before splicing them together to create \( N \) audio files. For each generated audio, a room impulse response was convolved, mixing \( N \) long recordings with a noise signal set at a random signal-to-noise ratio. The silence duration between speech segments was determined by an exponential distribution with a parameter of \( \beta \). Consequently, \( \beta \) can be utilized to control the speaker overlap rate of the generated audio, where a larger \( \beta \) value results in lower speaker overlap [16].
The corpus used for the simulation comprises both Switchboard-2 (Phases I, II, and III), Switchboard Cellular (Parts 1 and 2) and NIST Speaker Recognition Evaluation (2004, 2005, 2006, and 2008), with all audio being telephone speech sampled at 8 kHz. In total, these corpora contain 6381 speakers, divided into 5743 speakers for the training set and 638 speakers for the test set. As there are no temporal annotations in these corpora, similar to [16], we utilized SAD based on a time-delayed neural network (TDNN) and the Kaldi speech toolkit [49] for extraction. Noise data were obtained from 37 background noise segments from the MUSAN corpus [50], while the Room Impulse Response (RIRs) data came from the simulated RIR dataset used in [51], with 10,000 records selected. SNR values were chosen from 10, 15, and 20 dB. We generated mixed audio datasets with two and three speakers separately using different $\beta$ values to control the speaker overlap rate of the created mixed audio [16]. Each speaker had 20–40 utterances, with detailed simulated data information presented in Table 1.

Table 1. Simulated dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Split</th>
<th>Spk</th>
<th>Mixtures $^1$</th>
<th>$\beta$</th>
<th>Overlap Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim2spk</td>
<td>train</td>
<td>2</td>
<td>100,000</td>
<td>2</td>
<td>34.2%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>2</td>
<td>500</td>
<td>2</td>
<td>34.4%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>2</td>
<td>500</td>
<td>3</td>
<td>27.2%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>2</td>
<td>500</td>
<td>5</td>
<td>19.1%</td>
</tr>
<tr>
<td>Sim3spk</td>
<td>train</td>
<td>3</td>
<td>100,000</td>
<td>5</td>
<td>34.2%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>3</td>
<td>500</td>
<td>5</td>
<td>34.6%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>3</td>
<td>500</td>
<td>7</td>
<td>27.4%</td>
</tr>
<tr>
<td></td>
<td>test</td>
<td>3</td>
<td>500</td>
<td>11</td>
<td>19.2%</td>
</tr>
</tbody>
</table>

$^1$ Mixtures: total number of audios.

We also select audios containing two and three speakers from the real telephone conversation dataset CALLHOME to construct new datasets, referred as CALLHOME-2spk and CALLHOME-3spk, respectively, and employed Kaldi’s script to split them into two subsets. For CALLHOME-2spk, these subsets contained a fine-tuning set of 155 audios and a test set of 148 recordings, whereas for CALLHOME-3spk, the subsets had a fine-tuning set of 61 audios and a test set of 74 recordings. Further details are available in Table 2.

Table 2. CALLHOME dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Split</th>
<th>Spk</th>
<th>Mixtures $^1$</th>
<th>Overlap Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALLHOME-2spk</td>
<td>Part 1</td>
<td>2</td>
<td>155</td>
<td>14.0%</td>
</tr>
<tr>
<td></td>
<td>Part 2</td>
<td>2</td>
<td>148</td>
<td>13.1%</td>
</tr>
<tr>
<td>CALLHOME-3spk</td>
<td>Part 1</td>
<td>3</td>
<td>61</td>
<td>19.2%</td>
</tr>
<tr>
<td></td>
<td>Part 2</td>
<td>3</td>
<td>74</td>
<td>17.0%</td>
</tr>
</tbody>
</table>

$^1$ Mixtures: total number of audios.

4.2. Experimental Setup

For audio features, we follow the configuration of the original paper, utilizing a 23-dimensional MFCC for SA-EEND [11], and 80-dimensional Fbank features for CB-EEND [12] and our BSAC-EEND, with a frame length of 25 ms and a frame shift of 10 ms. Regarding SpecAugment, we employed two frequency masks and two times masks, with each frequency mask capable of masking up to two consecutive frequency channels and each time mask having at most 120 consecutive time steps.

In terms of subsampling, we adopted three methods: frame stacking subsampling, as used in the original SA-EEND [11]; DSC-based convolutional subsampling, as used in the original CB-EEND [12]; and our proposed BSCConv-based convolutional subsampling. For frame-stacked subsampling, we concatenated each feature with the seven preceding and following frames. For DSC-based convolutional subsampling, in line with [12], we
implemented two layers of two-dimensional (2D) depthwise separable convolutions [18]. For BSConv-based convolutional subsampling, we utilized CS-BSCU and CS-BSCS convolutional subsampling methods. These techniques subsampled the temporal domain of input features by a factor of 10 (equivalent to 100 milliseconds). Following [12], we set the convolution kernel parameters for all convolution subsampling as follows: \((3, 3), (7, 7)\). For 23-dimensional acoustic features, the step sizes were \((2, 1), (5, 1)\), and for the 80-dimensional acoustic features, the step sizes were \((2, 2), (5, 2)\).

For SA-EEND, we employed four transformer encoders, with each encoder containing four attention heads. Each attention head generated a 256-dimensional frame-by-frame embedding vector, and the feedforward layer consisted of 1024 internal units. Positional encoding was not utilized. For CB-EEND and BSAC-EEND, we used four conformer encoders, with each encoder containing four attention heads. Each attention head generated a 256-dimensional frame-by-frame embedding vector, and the feedforward layer comprised 1024 internal units. Similarly, positional encoding was not used, and the convolution module’s kernel size was set to 31. Furthermore, we applied the designed BSConv-based convolutional subsampling and MFA to CB-EEND for comparison.

In the training process with simulated data, we utilized the Adam optimizer [52] to update the neural network and the Noam scheduler [15] to adjust the learning rate, employing 100,000 warm-up steps. The Adam optimizer with a fixed learning rate was also employed for adaptation on real datasets. For efficient batching during training, audio was divided into 50 s segments when working with simulated data, and 200 s segments for the adaptation set. The training batch size was set to 64. During the inference stage, the entire audio was processed by the network without segmentation. We assessed the performance using the DER, which is defined by \(T_{\text{MI}}, T_{\text{FA}}, T_{\text{CF}}\), and \(T_{\text{Speech}}\) as the total speech duration, missed speech duration, false positive speech duration, and speaker confusion duration, respectively.

\[
\text{DER} = \frac{T_{\text{MI}} + T_{\text{FA}} + T_{\text{CF}}}{T_{\text{Speech}}}
\]  

we assessed performance on both the simulated dataset and the standard CALLHOME dataset, employing a collar tolerance of 0.25 s while comparing the hypothesized and reference speaker boundaries. Notably, speaker overlap was not excluded from the evaluation.

5. Result and Analysis

5.1. Improved Convolutional Subsampling

To investigate the performance of the proposed enhanced convolutional subsampling layer, we incorporated two convolutional subsampling structures depicted in Figure 3, substituting the respective elements in CB-EEND. These improvements were assessed on simulated datasets featuring two and three speakers, as well as real datasets. For the simulated data evaluation, the training sets of Sim2spk and Sim3spk were employed to train the EEND model, which was subsequently tested on their corresponding test sets (outlined in Table 1). Regarding the real dataset evaluation, the model was adapted using part 1 and was tested on part 2.

Table 3 presents the diarization error rate (DER) evaluation results. From the first three rows, it is evident that CB-EEND outperformed both the cluster-based cascade system and transformer-based EEND in terms of DER. Examining the last three rows, we observe that compared to the original CB-EEND, both CS-BSCU and CS-BSCS achieved significant improvements in DER when replacing the DSC-based convolutional subsampling layer. Specifically, the CS-BSCs layer resulted in 8.2% (\(\beta = 2\)), 15% (\(\beta = 3\)), and 14.9% (\(\beta = 5\)) relative improvements on the Sim2spk test set, a 3.6% improvement on the CALLHOME-2spk Part 2, and 17.1% (\(\beta = 5\)), 23.8% (\(\beta = 7\)), and 27.4% (\(\beta = 11\)) improvements on the Sim3spk test set. Additionally, a 1.2% relative improvement was observed on the CALLHOME-3spk Part 2.
Table 3. Improved convolutional subsampling evaluation results (DER).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sim2spk</th>
<th>CALLHOME-2spk</th>
<th>Sim3spk</th>
<th>CALLHOME-3spk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overlap Ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>34.4%</td>
<td>27.3%</td>
<td>19.1%</td>
<td>13.1%</td>
</tr>
<tr>
<td>x-vector (TDNN) + AHC [47]</td>
<td>28.77</td>
<td>24.46</td>
<td>19.78</td>
<td>11.53</td>
</tr>
<tr>
<td>CB-EEND</td>
<td>3.88</td>
<td>3.59</td>
<td>3.54</td>
<td>9.32</td>
</tr>
<tr>
<td>CB-EEND + CS-BSCS</td>
<td>3.56</td>
<td>3.05</td>
<td>3.01</td>
<td>8.98</td>
</tr>
<tr>
<td>CB-EEND + CS-BSCU</td>
<td>3.36</td>
<td>3.28</td>
<td>3.43</td>
<td>9.13</td>
</tr>
</tbody>
</table>

The CS-BSCU layer yielded 13.4% ($\beta = 2$), 8.6% ($\beta = 3$), and 3.1% ($\beta = 5$) relative improvements on the Sim2spk test set, and a 3.1% improvement on CALLHOME-2spk Part 2. Furthermore, 20.7% ($\beta = 5$), 21.1% ($\beta = 7$), and 28.7% ($\beta = 11$) relative improvements were achieved on the Sim3spk test set, along with a 1.4% improvement on the CALLHOME-3spk Part 2.

Overall, within CB-EEND, the CS-BSCS exhibits slightly better performance for two-speaker datasets and slightly inferior results for three-speaker datasets in comparison to the CS-BSCU layer. As discussed in Section 3.1, the convolutional subsampling based on BSConv outperforms the DSC-based approach, primarily due to its ability to reduce correlation within the convolution kernel more effectively. This facilitates the retention of more relevant information during the downsampling process.

5.2. MFA

We evaluated the performance of the MFA structure, as outlined in Section 3.3, within both SA-EEND and CB-EEND frameworks. Employing the experimental settings discussed in Section 4.1, we implemented the MFA structure in CB-EEND and assessed its performance on simulated and real datasets containing two or three speakers. Similar to Section 5.1, for the simulated data evaluation, we used the Sim2spk and Sim3spk training sets to train the EEND model and tested it on the respective test sets provided in Table 1. For real data evaluation, we utilized Part 1 of the CALLHOME dataset for adaptation and conducted testing on Part 2.

Table 4 presents the diarization error rate (DER) evaluation results. Compared with CB-EEND, the incorporation of the MFA structure yielded noticeable improvements in both simulated and real datasets. In the Sim2spk test set, relative improvements of 8.7% ($\beta = 2$), 8.3% ($\beta = 3$), and 4.5% ($\beta = 5$) were achieved, whereas a 3.3% improvement was observed in the CALLHOME-2spk set. Furthermore, the model attained 12.3% ($\beta = 5$), 13.3% ($\beta = 7$), and 17.3% ($\beta = 11$) relative improvements in the Sim3spk test set. Additionally, a 2.3% relative improvement was observed on CALLHOME-3spk Part 2.

Table 4. The evaluation results of the MFA structure (DER).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sim2spk</th>
<th>CALLHOME-2spk</th>
<th>Sim3spk</th>
<th>CALLHOME-3spk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overlap Ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>34.4%</td>
<td>27.3%</td>
<td>19.1%</td>
<td>13.1%</td>
</tr>
<tr>
<td>CB-EEND</td>
<td>3.88</td>
<td>3.59</td>
<td>3.54</td>
<td>9.32</td>
</tr>
<tr>
<td>CB-EEND + MFA</td>
<td>3.54</td>
<td>3.29</td>
<td>3.38</td>
<td>9.01</td>
</tr>
</tbody>
</table>

To better illustrate the impact of the MFA structure, we randomly selected a 50 s speech input from Sim2spk ($\beta = 2$) and reduced the high-dimensional features obtained prior to the linear layer to a two-dimensional space using principal component analysis (PCA). Figure 4 displays the visualization results for features extracted by SA-EEND, CB-EEND, and CB-EEND+MFA. Evidently, both CB-EEND and CB-EEND+MFA outperformed SA-EEND when discerning between silence frames and single-speaker speech frames, with CB-EEND+MFA proving superior in differentiating overlapping speaker frames. This
demonstrates that the model benefits substantially from this multiscale representation convergence.

Figure 4. Visualization of model extraction features.

5.3. BSAC-EEND

In this section, we evaluate the proposed BSAC-EEND model, employing the two convolutional subsampling layer structures depicted in Figure 3. The evaluation dataset includes both two- and three-speaker datasets, incorporating simulated and real datasets for training and testing. The model’s training and testing processes are consistent with those described in the previous two sections.

The evaluation results can be observed in Table 5. When compared to CB-EEND, BSAC-EEND (CS-BSCS) achieved relative improvements of 20.8% ($\beta = 2$), 16.1% ($\beta = 3$), and 24.8% ($\beta = 5$) on the two-speaker dataset and a relative improvement of 7.8% on CALLHOME-2spk. For the three-speaker dataset, it showed relative improvements of 2.5% ($\beta = 5$) and 3.5% ($\beta = 7$). BSAC-EEND (CS-BSCU) demonstrated relative improvements of 16.2% ($\beta = 2$), 8.3% ($\beta = 3$), and 15.8% ($\beta = 5$) on the two-speaker dataset, with a 6.1% relative improvement on CALLHOME-2spk. On the three-speaker dataset, relative improvements of 15.2% ($\beta = 5$), 10.8% ($\beta = 7$), and 21.7% ($\beta = 11$) were recorded, with a 3.79% relative improvement on CALLHOME-3spk.

Table 5. Evaluation results for BSAC-EEND (DER).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sim2spk</th>
<th>CALLHOME-2spk</th>
<th>Sim3spk</th>
<th>CALLHOME-3spk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overlap Ratio</td>
<td>34.4%</td>
<td>27.3%</td>
<td>19.1%</td>
<td>13.1%</td>
</tr>
<tr>
<td>CB-EEND</td>
<td>3.88</td>
<td>3.59</td>
<td>3.54</td>
<td>9.32</td>
</tr>
<tr>
<td>BSAC-EEND (CS-BSCS)</td>
<td>3.07</td>
<td>3.01</td>
<td>2.66</td>
<td>8.59</td>
</tr>
<tr>
<td>BSAC-EEND (CS-BSCU)</td>
<td>3.25</td>
<td>3.29</td>
<td>2.98</td>
<td>8.75</td>
</tr>
<tr>
<td>EEND-EDA (Chronol) [47]</td>
<td>3.07</td>
<td>2.74</td>
<td>3.04</td>
<td>8.24</td>
</tr>
<tr>
<td>EEND-EDA (Shuffled) [47]</td>
<td>2.69</td>
<td>2.44</td>
<td>2.60</td>
<td>8.07</td>
</tr>
</tbody>
</table>

These findings indicate that BSAC-EEND achieves an overall DER improvement compared to SA-EEND and CB-EEND, with BSAC-EEND (CS-BSCS) performing better on the two-speaker dataset and BSAC-EEND (CS-BSCU) showing superior results on the three-speaker dataset. This outcome aligns with the analysis presented in Section 5.1. We conducted a comparative evaluation with the cutting-edge end-to-end speaker diarization system, EEND-EDA [48]. As reported in [48], EEND-EDA exhibits strong performance with fewer speakers, but there is significant room for improvement when working with more complex speaker datasets. Comparing the results in the table, our system demonstrates slightly inferior performance to EEND-EDA for the two-speaker dataset, but outperforms...
it in the more challenging three-speaker dataset. This substantial improvement suggests that our approach is better equipped to handle intricate audio information.

6. Conclusions

In this study, we introduce a conformer-based EEND system with an MFA structure and a novel convolutional subsampling layer. Our experimental results demonstrate that the EEND system benefits from the convergence and fusion of multi-scale representations facilitated by the MFA structure and the innovative convolutional subsampling layer. Although our system surpasses SA-EEND and CB-EEND in terms of overall performance, we observed that even after fine-tuning on real datasets, the performance of the new EEND system on the CALLHOME dataset appears relatively restrained, a dilemma potentially attributable to the domain mismatch problem existing between simulated training data and actual test data [12,53]. As a future direction, we plan to incorporate transfer learning techniques to enhance the EEND system’s training process, enabling better performance on real test data. Within the context of system architecture, existing literature [54] has suggested that quantum convolutional neural networks may enhance speech signal processing performance by extracting more representative speech features. Thus, employing quantum convolutional neural networks for future downsampling processes could be a potential approach. Moreover, given the importance of privacy protection, particularly in audio information modeling [55], it emerges as a potential area warranting further research focus in the future.

Author Contributions: Conceptualization, software, writing—original draft, X.J.; validation, visualization, Y.C.; supervision, writing—review and editing, D.Q. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 62171470), the Central Plains Science and Technology Innovation Leading Talent Project of Henan Province (No. 234200510019), and the General Project of the Henan Provincial Natural Science Foundation of China (No. 232300421240).

Data Availability Statement: The simulated datasets used in the current study are not publicly available as these datasets can be generated by the readers themselves based on the simulated settings in Section 4, and the real datasets used can be purchased online.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

- EEND: End-to-end neural diarization
- DSC: Depthwise separable convolution
- BSConv: Blueprint separable convolution
- MFA: Multi-scale feature aggregation
- BSAC-EEND: Blueprint separable subsampling and aggregate feature conformer-based EEND
- ASR: Automatic speech recognition
- SA-EEND: Self-Attentitive End-to-End Neural Diarization
- BSConv-U: Unconstrained BSConv
- BSConv-S: Subspace BSConv
- MFA-Conformer: Multi-scale feature aggregation conformer
- SAD: Speech activity detection
- AHC: Aggregated hierarchical clustering
- VBx: Bayesian HMM clustering of x-vector sequences
- BLSTM: Bidirectional long short-term memory
- CB-EEND: Conformer-based EEND
- CS-BSCU: Convolutional subsampling based on BSConv-U
References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.