NAVER CLOVA SUBMISSION TO THE THIRD DIHARD CHALLENGE

Hee-Soo Heo[1], *Jee-weon Jung*[1,2], *Youngki Kwon*[1], *You Jin Kim*[1],
Jaesung Huh[3], *Joon Son Chung*[1], *Bong-Jin Lee*[1]

[1]Naver Corporation, South Korea
[2]School of Computer Science, University of Seoul, South Korea
[3]Visual Geometry Group, Department of Engineering Science, University of Oxford, UK

ABSTRACT

This report describes the NAVER CLOVA speaker diarization system for the third DIHARD challenge. Our system comprises the following five subsystems: end-point detection, overlapped speech detection, speaker embedding extraction, feature enhancement, and clustering. Its process pipeline has two improvements over the conventional diarization systems: feature enhancement and overlapped speech detection. For feature enhancement, our proposed approach first adopts an utterance-wise autoencoder that reduces the dimensionality of extracted speaker embeddings. Then, we apply a self-attention mechanism in which we refer to as the attention-based aggregation. We aim to adapt and enhance the speaker representation for clustering using these two techniques. Also, variants of CRNN based overlapped speech detection systems, trained as a three-class classifier, and their ensemble are explored to further reduce the missed detection of overlapped speech regions. The submitted system achieves a diarization error rate of 14.96% and 15.40% for the development and the evaluation datasets of `DIHARDIII_Task1_CORE` track, which ranks the 3^{rd} place.

***Index Terms*—** Speaker diarization, feature enhancement, overlapped speech detection.

1. INTRODUCTION

The overall framework of our speaker diarization system is illustrated in Figure 1. It comprises end-point detection (EPD), overlapped speech detection (OSD), speaker embedding extraction, feature enhancement, and clustering modules. In the following sections, we introduce more details of each module.

2. END-POINT DETECTION

We use the EPD module presented in the baseline of DIHARDIII challenge for track2 [1, 2]. In particular, the EPD module smooth the time-delay neural network-based speech activity detection result with a hidden Markov model following Kaldi Aspire recipe [3].

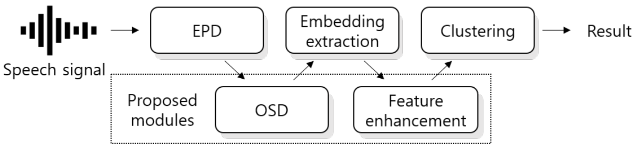

Fig. 1: Process pipeline of the NAVER CLOVA speaker diarization system.

3. OVERLAPPED SPEECH DETECTION

The OSD system detects the onsets and the offsets of segments that contain more than one speaker's speech. Our OSD system has three features, compared to the conventional systems. First, we train the model as a three-class classifier, namely non-speech, single speaker speech, and overlapped speech. In the test phase, we use the output layer's node that indicates overlapped speech. Second, we augment the training dataset by adding another speaker's segment in the middle of a single speaker segment similar to that in [4]. This augmentation is performed to increase the ratio of overlapped speech, balancing ratios between different classes. Last, we use a convolutional recurrent neural network (CRNN) architecture instead of recurrent neural network-based existing systems inspired by the recent success of CRNN architectures in sound event detection [5].

For more reliable results, we use a score-level ensemble of three variants: 2D CRNN model with squeeze and excitation (SE) [6], 2D CRNN model without SE, and 1D CRNN model without SE. The three model variants share a 128-dimensional Mel-spectrogram as the input feature. The architectures of each variant are shown in Table 1. Figure 2 shows the performances of our OSD modules on the development set. We set the threshold to meet a precision of 0.8, which is most effective in terms of the diarization error rate (DER) on the development set.

Based on the reported threats of overlapped speech in speaker diarization [4], we modify our process pipeline by incorporating the proposed OSD module. We first extract speaker embeddings only from single speaker segments and

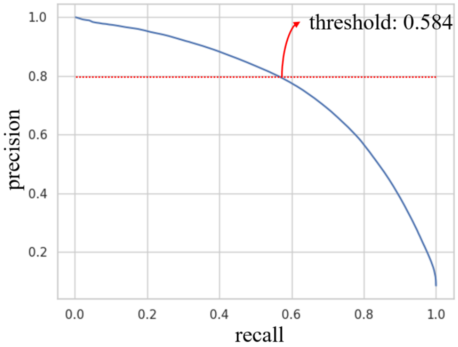

Fig. 2: Precision-recall curve illustrating the performance of the ensemble of three OSD variants where a thresholds that give a precision 0f 0.8 for the DIHARDIII development set is chosen.

perform speaker clustering. Then, for each overlapped segments, we predict two labels based on centroids of adjacent clusters. Concretely, we find top-2 centroids showing the highest similarity and assign the corresponding labels for each embedding of overlapped speech where each centroid is calculated by averaging the embeddings of the corresponding cluster.

4. SPEAKER EMBEDDING EXTRACTION

For the speaker embedding extraction, we train the ResNet34SEV2 architecture in voxceleb_trainer[1] with a few modifications. We use the development set of both VoxCeleb1 and 2 datasets to train the model. The number of filters in the first convolutional layer is configured to 64. Average pooling is performed instead of attentive statistics pooling after the last convolutional operation. We extract a 256 dimensional speaker embedding from each segment with a window of 1.25 second width, and 0.125 second shift.

5. FEATURE ENHANCEMENT

Feature enhancement is common in the machine learning field, however, it has not been explored for speaker diarization to the best of our knowledge. In our analysis, speaker embeddings in speaker diarization require discriminative power for only a small number of speakers (e.g., four or less). This is in contrast to speaker embeddings of speaker identification or verification which demands discrimination of thousand of speakers. Based on this analysis, we presume that the diarization system can be improved by projecting speaker embeddings to another representation space where a small number of speakers from an identical session are well discriminated.

Table 1: Architectures of CRNN-based variants for OSD. Each convolutional block is composed of two convolutional layers with batch normalization layers and an average pooling layer. There is no stride for all convolutional layers. Therefore the output shapes of convolutional blocks are determined by the pooling size. **L**: length of the input sequence in frames, **bi-GRU**: bidirectional gated recurrent unit.

Layer	Kernel size	Output shape
	1D CRNN w/o SE	
1D Conv Block	$128 \times 3 \times 128$	$L \times 128$
1D Conv Block	$128 \times 3 \times 196$	$L/2 \times 196$
1D Conv Block	$196 \times 3 \times 256$	$L/6 \times 256$
bi-GRU	256×512	$L/6 \times 1024$
	2D CRNN w/o SE	
2D Conv Block	$1 \times 3 \times 3 \times 32$	$L/2 \times 128 \times 32$
2D Conv Block	$32 \times 3 \times 3 \times 64$	$L/6 \times 64 \times 64$
2D Conv Block	$64 \times 3 \times 3 \times 128$	$L/6 \times 32 \times 128$
Avg pooling	-	$L/6 \times 128$
bi-GRU	128×256	$L/6 \times 512$
	2D CRNN w/ SE	
2D Conv Block	$1 \times 3 \times 3 \times 32$	$L/2 \times 128 \times 32$
2D Conv Block	$32 \times 3 \times 3 \times 64$	$L/6 \times 64 \times 64$
2D Conv Block	$64 \times 3 \times 3 \times 128$	$L/6 \times 32 \times 128$
Avg pooling	-	$L/6 \times 128$
bi-GRU	128×256	$L/6 \times 512$

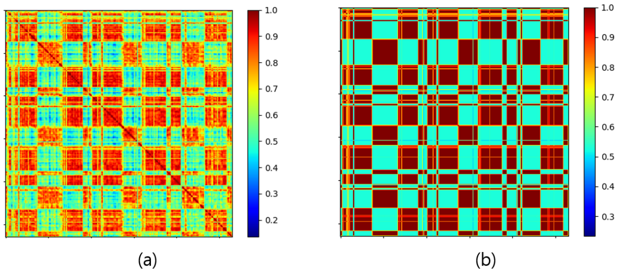

Fig. 3: Effect of the attention-based aggregation technique on the affinity matrix. We calculate the affinity matrix using one sample in development set to compare the before (a) and after (b) applying the attention-based aggregation technique. The results show that the proposed technique act like a refinement process that removes the most of noises on the affinity matrix.

To achieve this goal, we first reduce the dimensionality of speaker embeddings using an autoencoder, that is trained for each session during run-time. The autoencoder comprises two layers, one for the encoder and the other for the decoder. For the encoder layer, we apply max feature-map activation [7]. Using the trained autoencoder, the 256-dimensional embedding vectors are projected into a 20-dimensional enhanced representations. Note that in dimensionality reduction, the autoencoder is designed to learn reconstruction of only one session after random initialization. In particular, we train the au-

[1] https://github.com/clovaai/voxceleb_trainer

toencoder by 200 epochs using Adam optimizer with a 0.001 learning rate for each session.

After dimensionality reduction, we aggregate embedding vectors using a self-attention mechanism, described in Algorithm 1. We calculate the attention map for each embedding using the softmax function and update embeddings iteratively based on the derived attention maps. Two hyper-parameters needs to be configured to apply the proposed attention-based aggregation: number of repetitions and temperature value before softmax function. We fix these two values as 5 and 15, respectively. Figure 3 shows the effect of the aggregation technique on the affinity matrix.

6. CLUSTERING MODULE

Based on embeddings enhanced by the proposed modules, we perform spectral clustering to estimate the speaker labels [8, 9]. First, we calculate the affinity matrix using cosine similarity. Note that we do not apply any additional refinement process on this matrix. Then, eigen-values and eigen-vectors are calculated by applying eigen-decomposition to the affinity matrix. To determine the number of clusters, eigen-values greater than 20 are counted. Finally, we perform k-means clustering on the spectral embeddings, which is a set of eigenvectors corresponding to the largest eigen-values, to estimate final cluster labels.

7. DATASETS

Since our diarization system is not an end-to-end system, achieving the best performance in each module does not always guarantee the lowest DER. Therefore, we empirically explore and tune the performance of our diarization system, a combination of various modules, using the datasets listed in Table 2.

8. RESULTS

Table 3 demonstrates performances of various systems on the `DIHARDIII_Task1_CORE` track, using the DIHARDIII development set. First, we find that the proposed feature enhancement technique significantly increases the performance compared to the baseline. Additional performance improvements are achieved using the output of the OSD module. We also find that the proposed OSD and feature enhancement modules can be effectively applied to conventional clustering methods such as agglomerative hierarchical clustering. We submitted the system that achieved 14.97% DER on the development set as the primary system. In addition, the submitted system showed an average time efficiency of 0.014 response time for the entire development set. The time efficiency was measured using one 24GB NVIDIA Tesla P40 GPU, and the maximum memory usage was 10.49GB. Finally, Table 4 shows the results from the leaderboard of the DIHARDIII challenge.

Algorithm 1 Attention-based aggregation

```
Input: Speaker embeddings $\mathbf{X}$ of size $L \times 20$
Hyper-parameters: Number of repetition $N$, Temperature value $\tau$
for $iteration = 1, 2, \ldots, N$ do
    Construct affinity matrix $\mathbf{A}|\mathbf{A}_{i,j}$=cos($\mathbf{X}_i$,$\mathbf{X}_j$)
    $\mathbf{A}$ = softmax($\mathbf{A}$ $*\tau$)
    $\mathbf{X}$ = dot($\mathbf{A}$, $\mathbf{X}$)
end for
```

Table 2: Data configurations for building each module of the NAVER CLOVA speaker diarization system.

Module	Data & Set	Label
Embedding extractor	VoxCeleb dev VoxCeleb 2 dev MUSAN [10] simulated rir [11]	speaker label
OSD module	AMI DIHARD I dev DIHARD II dev VoxConverse dev MUSAN	RTTM
EPD module (only for track2)	DIHARD III dev	speech activity
Hyper-parameter tuning	DIHARD III dev	RTTM

Table 3: Performances of various systems on development set in diarization error rate (DER) for task1. **FE**: feature enhancement, **OSD**: overlapped speech detection.

	Clustering method	DER (%)
Our baseline	AHC	21.41
	Spectral clustering	25.23
W/ FE	AHC	17.68
	Spectral clustering	16.84
W/ FE & OSD	AHC	16.45
	Spectral clustering	14.97

Table 4: Performances of the submitted system on evaluation set in diarization error rate (DER). **FE**: feature enhancement, **OSD**: overlapped speech detection.

	Part	DER (%)	
		track1	track2
Primary system W/ FE & OSD	core	15.40	24.31
	full	13.95	21.86

9. REFERENCES

[1] Prachi Singh, Harsha Vardhan, Sriram Ganapathy, and Ahilan Kanagasundaram, "Leap diarization system for the second dihard challenge," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019): Crossroads of Speech and Language*. International Speech Communication Association, 2019, pp. 983–987.

[2] Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman, "The third dihard diarization challenge," *arXiv preprint arXiv:2012.01477*, 2020.

[3] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[4] Latane Bullock, Herve Bredin, and Leibny Paola Garcia-Perera, "Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection," in *ICASSP*. 2020, vol. 2020-May, pp. 7114–7118, IEEE.

[5] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark D. Plumbley, "Polyphonic Sound Event Detection and Localization using a Two-Stage Strategy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. 2019, pp. 30–34, New York University.

[6] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-Excitation Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[7] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[8] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[9] Huazhong Ning, Ming Liu, Hao Tang, and Thomas S Huang, "A spectral clustering approach to speaker diarization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[10] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[11] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.