GREAT-SHU Speaker Diarization System for DIHARD III Challenge

Zhiyong Chen, Zongze Ren, Runze Ma, Bo Wu, Shugong Xu
Shanghai Institute for Advanced Communication and Data Science
Shanghai University
Shanghai, China

***Abstract*—In this paper, we present the submitted system for the third DIHARD Speech Diarization Challenge from the GREAT-SHU team. Our diarization system includes multiple modules, namely voice activity detection (VAD), segmentation, speaker embedding extraction, similarity scoring, matrix refinement, and clustering. For each module, we explore different techniques to enhance performance. Our final submission employs the oracle VAD, uniform-refined segmentation, the Deep ECAPA-TDNN based speaker embedding, the cosine similarity based similarity scoring and spectral clustering. We did not implement resegmentation stage and overlap detection. Our proposed system achieves 22% DER in Track1 and 39% DER in Track2. We believe that the diarization task is still challenging.**

***Index Terms*—DIHARD, VAD, speaker embedding, similarity scoring, clustering**

I. System Description

Speaker diarization is the task of determining "who spoke when" in an audio file that usually contains an unknown number of speakers with variable speech duration. It has a wide range of applications such as telephone calls, meeting recordings and broadcast interviews. Diarization can also serve as the frontend of automatic speech recognition (ASR) to improve the transcription performance in multi-speaker conversations.

A. Main system description

Our work is basically a reproduction of the system [1] proposed by DKU in DIHARD2 challenge, including following sections.

- Voice activity detection (VAD): VAD detects speech in the audio signals and removes the non-speech regions to reduce computation. We used WebRTC as our VAD system for Track2.
- Segmentation: The segmentation step splits speech into multiple speaker-homogeneous segments. We use uniform segmentation with overlap, the segmentation length is set to 2 seconds.
- Speaker embedding extraction: After segmentation, short segments are mappedd into the speaker sub-space and generate fixed-dimensional speaker embeddings. Here we use newly-proposed ECAPA-TDNN system for speaker embedding extraction.
- Similarity measurement: Similarity scores between any two speaker embeddings in the same audio are computed and later used in the clustering step. Popular techniques includes cosine similarity, probabilistic linear discriminant analysis (PLDA). We only use cosine similarity in our system.
- Clustering: Clustering algorithms like K-means, agglomerative hierarchical clustering (AHC) and spectral clustering assign segments with high similarity scores to the same cluster. We use spectral clustering in our system.

B. Data

The experiments are conducted with the development set of Voxceleb 2, which contains 1,092,009 utterances from 5,994 speakers. For evaluation, the evaluation set of DIHARD3 are used. We report the experimental results on DIHARD3 leader board for both core and full condition.

We perform online data augmentation [1] with MUSAN dataset [1]. The noise type includes ambient noise, music, television, and babble noise for the background additive noise. The television noise is generated with one music file and one speech file. The babble noise is constructed by mixing three to eight speech files into one. For the reverberation, the convolution operation is performed with 40,000 simulated room impulse responses (RIR) in MUSAN. We only use RIRs from small and medium rooms. Moreover, we adopt the speed perturbation using sox to increase the speaker number. The strategy also has a successful application in speech and speaker recognition tasks. We speed up or down each utterance by 0.9 or 1.1 times, and the utterances with different speeds are considered from new speakers. We add inverse spectral augmentation, therefore 20% of the frequency bins are randomly masked with zeros. Finally, we balanced sampled utterances from 5,994×3 = 17,982 speakers in each epoch.

C. Similarity Matrix Refinement

We perform similarity matrix refinement on the original matrix. As shown in Fig 1, we set boundary values of the matrix intensity and normalized the matrix intensity within that boundary. The sparse activation in the matrix is therefore surpressed.

II. Experiments

Our system results is shown in Tab I. The matrix refinement show its effectiveness on the JER matric. Though the overall performance of our proposed system is limited in this first

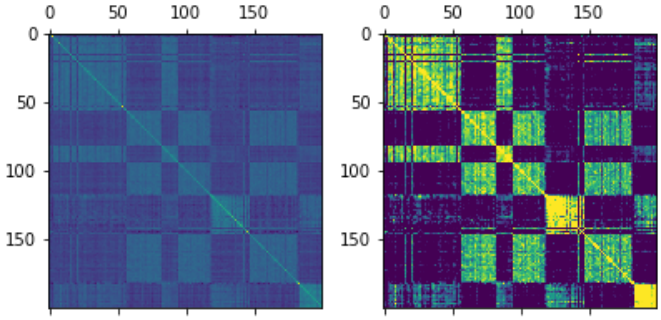

Fig. 1. Left: original similarity matrix. Right: refined similarity matrix

TABLE I
PRIMARY RESULTS ON DIHARD3 FULL CONDITION.

System	DER%	JER%
Track1 original	22.630	47.280
Track1 refined	**23.470**	**39.780**
Track2 refined	39.580	55.080

attmept of the speaker diarization task, our research group will continuely focus on this research field and try to contribute in the future.

REFERENCES

[1] Wang, Weiqing and Cai, Danwei and Qin, Xiaoyi and Li, Ming, "The DKU-DukeECE Systems for VoxCeleb Speaker Recognition Challenge 2020,".
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.