Adaptive Speaker Identification and Speaker Diarization System in dh_DiaboLIIc Team Submission for DIHARD III Challenge

Wenda Chen[*], *Sangeeta Ghangam*[*]

USA

wendacyema@gmail.com, sghangam2010@gmail.com

Abstract

The DIHARD III challenge introduced many new challenges – segmentation breaks $>$ 200ms, different types of human speech sounds in the complete dataset and an average segment length less than 2s, amongst others. In order to successfully perform speaker diarization on the DIHARD III dataset, the dh_DiaboLIIc team used different approaches for segmentation, embedding and re-segmentation resulting in a multi-system ensemble submission. This workshop abstract submission will focus on the speaker embedding models and re-segmentation pipelines utilized as part of the challenge submission. These techniques are shown to adapt the general models and systems to DIHARD III development data and improve the results in the local domains.

1. Introduction

Speaker embeddings are important for diarization and clustering. DIHARD III data is very challenging with about 1.9s average segment length and limited enrollment data for each speaker. In the previous work [1], it has been shown that shortened speech segments will lead to less accurate embeddings. The x-vector with attention system was trained with additive margin softmax loss (AM-Softmax) and the intra-distance of the embeddings from the same speaker but with different lengths and noises, called IRL approach [1].

Speaker diariation is very important for real-time conversational audio data processing. In this research, the first system included the x-vector models first trained using Voxceleb dataset [2] and further fine-tuned with the DIHARD III development data using manual segmentations. This was integrated as part of the Pyannote pipeline [3] to generate one set of the results. In the second system, Pyannote pretrained embeddings using Voxceleb were converted into Kaldi [4] segment and ark files for processing by the the BUT VBx resegmentation system [5] with fine-tuned parameters for diarization evaluations.

2. Algorithms and System Description

This section summarizes the key algorithms we used in the system from embeddings to diarization. For the algorithms and results sections, we present both the embedding generation and the system level diarization processes.

2.1. Speaker Identification Model Adaptation

The speaker identification model was first trained on Voxceleb data. During adaptation, the x-vector model was applied on our VAD based segmentation of the audio files in DIHARD III and used to generate the 256 dimensional embedding for each segment. The 40 dimensional acoustic feature sequence is passed the 8 layers of x-vector every 25ms with 10ms shift. The audio segment labels in the DIHARD III development data were used to learn the target embedding output with AM-Softmax loss.

2.2. Pyannote Embeddings and VBx Diarization Pipeline

The pre-trained pyannote speaker embedding model (emb_voxceleb) was used as the starting point for the second system. This model was trained with a segment duration up to 4s which was critical for ensuring the accurate conversion of the embeddings into the Kaldi format for processing by the BUT VBx pipeline. For optimal results, the segment duration in VBx was extended to 2s with an overlap of 0.5s. In the VBx part of the BUT system, the interpolation alpha value was changed to 0.75, thereby altering the resulting PLDA mean value. Based on experiments with the DIHARD III development data, the Ploop value was fine-tuned to 0.40.

3. Results

3.1. Comparison of Embedding Models

The base models to generate embeddings were trained on Voxceleb dataset. We produced two attention-based x-vector models targeting the short-duration cases, one is general AM-Softmax model and the other one is IRL training of AM-Softmax model. The two models' test results on the different durations (1s, 2s, 4s, full length) of VOiCES dataset [6] are in Table 1. Since the average segment length of the speech segments in DIHARD III dataset is between 1s-4s, it will be important to explore the model performance on these durations. Table 1 shows that model 1 performs better on 1s-2s cases while model 2 performs better on 4s and above. This is further demonstrated in computing the average cosine distances of the embeddings from the same speaker in DIHARD III development data using the two models: model 1's embeddings have the lower distance 0.941 while model 2 has the higher distance 0.953, which agrees with the fact that the average segment length is around 2s.

Table 1: *Model performance (dev EER%) on varied lengths of test speech in VOiCES.*

Models	1s	2s	4s	Full
Model1: AM-Softmax	12.74	6.70	3.99	1.90
Model2: AM-Softmax-IRL	13.67	7.13	3.69	1.49

[*]Equal contribution, on behalf of dh_DiaboLIIc team, work done at USA site

3.2. Diarization Results

The preliminary diarization results on the baseline Pyannote system using the embeddings generated in Section 3.1 and the combined ensemble system with the second VBx system using DOVER-LAP [7] are summarized in Table 2. While the first system achieved good results, 4% further improvement in DER was observed with the system fusion on DIHARD III development data.

Table 2: *Combined ensemble result of the two systems from the scoring server vs results from baseline pyannote pipeline. An improvement of 4% in the DER results on DIHARD III development data.*

System	DER	JER
Pyannote-Baseline	28.70	47.88
Combined Ensemble	24.67	45.61

4. Discussion

This extended abstract summarizes two key components of our submitted system, namely speaker embedding generation and speaker diarization pipeline. The effectiveness of these approaches were demonstrated in the preliminary experiments. The final results submitted to the challenge included more components and techniques which resulted in better results on the evaluation sets of two tracks. Further details of the components will be reported in the final system descriptions.

5. References

[1] W. Chen, J. Huang, and T. Bocklet, "Length- and Noise-aware Training Techniques for Short-utterance Speaker Recognition," in *INTERSPEECH 2020*.

[2] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[3] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[5] L. F. N. DIEZ Sánchez Mireia, BURGET Lukáš and Jan, "Analysis of speaker diarization based on bayesian hmm with eigenvoice priors," in *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING vol. 28, no. 1, pp. 355-368. ISSN 2329-9290*, 2020.

[6] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (voices) corpus," arXiv 1804.05053, 2018.

[7] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," arXiv 2011.01997, 2020.