Hitachi-JHU System for the Third DIHARD Speech Diarization Challenge

Shota Horiguchi[1†], *Nelson Yalta*[1†], *Paola García*[2], *Yuki Takashima*[1], *Yawen Xue*[1], *Desh Raj*[2], *Zili Huang*[2], *Yusuke Fujita*[1], *Shinji Watanabe*[2], *Sanjeev Khudanpur*[2]

[1]Hitachi, Ltd.
[2]Johns Hopkins University

{shota.horiguchi.wk, nelson.yalta.wm}@hitachi.com, lgarci27@jhu.edu

Abstract

This paper describes Hitachi and JHU's joint efforts on the speaker diarization system for multi-talker monaural recordings evaluated on the Third DIHARD Speech Diarization Challenge. Our system focuses on the efficient combination of x-vector clustering and end-to-end diarization. With the state-of-the-art x-vector-based systems, end-to-end systems, and their combinations using EEND as post-processing and DOVER-Lap, we achieved DERs of $11.58\,\%$ in Track 1 and $16.94\,\%$ in Track 2 on evaluation set, respectively.

1. System description

Our system is based on both the conventional x-vector-based and end-to-end neural diarization (EEND) based methods.

1.1. X-vector-based system

The system consists of the following components:

Voice activity detector: We employed two voice activity detectors (VAD): SincNet-based VAD [1] and TDNN-based VAD. SincNet-based VAD learns to detect speech from the raw speech using a combination of a SincNet followed by BiLSTM layers and a fully connected layer. We trained the model using the DIHARD III dev set for 300 epochs. TDNN-based VAD consists of five-layer TDNN using statistics pooling. It was also trained on the DIHARD III dev set for 10 epochs. We also employed data augmentation using MUSAN [2] and simulated room impulse response for training the TDNN model. The final segmentation was obtained by averaging posteriors from the two models, followed by thresholding and median filtering.

Speaker embedding extractor: We employed TDNN [3, 4] and Res2Net [5] for extracting speaker embeddings.

The TDNN-based extractor uses 40-dimensional filterbanks, with a $25\,\mathrm{ms}$ window and $15\,\mathrm{ms}$ frame shift. These features are used for the embedding extraction as in [3]. The x-vector was trained using a TDNN with a $1.5\,\mathrm{s}$ window with frame shift of $0.25\,\mathrm{s}$. The TDNN extractor consists of four TDNN-ReLU layers each of them followed by a dense-ReLU. Then, two dense-ReLU layers are incorporated before a pooling layer; a final dense-ReLU is included from which 512-dimension embeddings are computed. A dense-softmax concludes this TDNN architecture [4].

The Res2Net-based extractor uses the default configuration described in [6]. The Res2Net uses 80 log-filterbank dimensions as input, a multi head-attentive pooling with attention heads set to 16 that learns to weight each frame, and additive angular margin Softmax [7] with margin of 0.1 and scale of 30 as a training criterion. For our experiments, we employed four extractors summarized in Table 1.

Table 1: *Res2Net-based x-vector extractors.*

	# of layers	Normalization	Compression	SpecAug
Res2Net-BN	23	BatchNorm	$\ln x$	
Res2Net-UN	23	UtteranceNorm	$\log_{10} x$	
Res2Net-BN-Large	50	BatchNorm	$\ln x$	
Res2Net-UN-Large	50	UtteranceNorm	$\log_{10} x$	✓

We employed the VoxCeleb 1 and VoxCeleb 2 sets [8] as training that provided 7323 speakers and over 1M of recordings. We augmented the data following a similar data augmentation as the Kaldi recipe for VoxCeleb[1]. Each audio recording is randomly chunked into subsegments of length between $2.0\,\mathrm{s}$ and $4.5\,\mathrm{s}$ that are feed into the models.

VBx: To eliminate the need for a tuned agglomerative hierarchical clustering (AHC) stopping threshold, we perform VBx-clustering after AHC [3]. The VBx-clustering is a simplified variational Bayes diarization. It follows an HMM, in which the state represents a speaker, and the state transitions correspond to speaker turns. The state distributions, or emission probabilities, are GMMs constrained by eigenvoice matrix. Each speaker has a probability of P_{loop} when the HMM ends up back in the same state. The initialization for this system is a PLDA model. For our experiments, this PLDA is the result of the interpolation of the VoxCeleb PLDA and the DIHARD III PLDA. Both PLDAs were centered and whitened using DIHARD III dev data.

For the TDNN-based system, the x-vectors were projected from 512 dim to 220 using a LDA, the PLDA interpolation regulated by an alpha was set to 0.50, and the value for P_{loop} to 0.80. For the Res2Net-based system, the 128-dimension embeddings were maintained, the alpha value was reduced to 0.1 and the P_{loop} to 0.80.

Overlap assignment: We used a similar approach to perform overlap detection as the one shown for the SincNet-based VAD, with the only difference that the classifier will distinguish between overlapping speech versus non-overlapping speech. We assigned the closest other speaker in the time axis as the second speaker for each detected frame.

For the TDNN-based system, the overlap assignment results were passed to the final system fusion in Section 1.4 (**System (1)**). For the res2net-based system, we applied the modified DOVER-Lap [9] (see Section 1.4) to combine the results of four res2net-based models (**System (2)**).

1.2. EEND-based system

For the EEND-based system, we employed EEND-EDA [11] to estimate diarization results and the number of speakers simulta-

[†]These authors contributed equally to this work.

[1]https://github.com/kaldi-asr/kaldi/blob/master/egs/voxceleb/v2

Table 2: *DERs / JERs (%) on Track 1 & 2.*

	Track 1 (w/ oracle VAD)				Track 2 (w/o oracle VAD)			
	Dev		Eval		Dev		Eval	
System	full	core	full	core	full	core	full	core
Baseline [10]	19.41 / 41.66	20.25 / 46.02	19.25 / 42.45	20.65 / 47.74	21.71 / 43.66	22.28 / 47.75	25.36 / 46.95	27.34 / 51.91
(1) TDNN-based x-vector + VBx + OvlAssign	13.87 / 32.73	14.88 / 36.72	15.65 / 33.71	18.20 / 38.42	17.61 / 36.03	18.64 / 39.92	21.47 / 37.83	24.58 / 42.02
(2) Res2Net-based x-vector + VBx + OvlAssign	14.04 / 34.29	15.18 / 38.80	15.81 / 35.53	18.47 / 40.47	17.26 / 37.17	18.39 / 41.56	21.37 / 39.59	24.64 / 44.49
(3) EEND-EDA	12.92 / 33.85	13.95 / 35.37	13.95 / 35.37	17.28 / 41.97	15.90 / 35.94	18.50 / 41.71	19.04 / 38.89	22.84 / 45.27
(4) SC-EEND	13.13 / 35.35	16.05 / 41.80	15.16 / 38.62	19.14 / 46.04	16.16 / 37.52	19.00 / 43.74	20.30 / 42.19	24.75 / 49.36
(5) TDNN-based x-vector + VBx + EENDasP	12.63 / 31.52	14.61 / 36.28	13.30 / 33.02	15.92 / 38.29	15.94 / 34.11	18.09 / 38.97	18.13 / 35.82	21.31 / 40.78
(6) DOVER-Lap of (1)(2)(3)(4)(5)	10.73 / 31.39	12.56 / 36.88	11.83 / 32.85	14.41 / 38.81	14.13 / 34.32	16.06 / 39.75	17.21 / 37.64	20.34 / 43.40
(7) EEND-EDA (SSA)	12.95 / 33.98	15.69 / 40.03	12.74 / 34.08	15.86 / 40.44	15.03 / 33.64	17.52 / 39.15	17.81 / 38.32	21.31 / 44.32
(8) TDNN-based x-vector + VBx + EENDasP (SSA)	12.54 / 31.32	14.55 / **36.11**	12.74 / **32.20**	15.34 / **37.50**	15.45 / 33.61	17.77 / **38.67**	17.60 / **35.16**	20.84 / **40.18**
(9) DOVER-Lap of (1)(2)(4)(7)(8)	**10.65** / **30.82**	**12.47** / 36.21	**11.58** / 32.37	**14.09** / 38.25	**13.85** / **33.41**	**15.81** / 38.77	**16.94** / 36.31	**20.01** / 41.78

neously. We followed the training strategy described in [11] but with larger number of epochs and mixtures of a larger number of speakers. We first trained the model using simulated two-speaker mixtures for 100 epochs, then fine-tuned it using simulated mixtures, each of which contains at most five speakers for 75 epochs, and finally adapted on the DIHARD III dev set for 200 epochs. The simulated data are based on the following corpora: Switchboard-2 (Phase I, II, III), Switchboard Cellular (Part 1, 2), NIST Speaker Recognition Evaluation (2004, 2005, 2006, 2008), and MUSAN corpus [2]. The training was based on log-Mel-filterbank based features extracted for each $100\,\mathrm{ms}$.

The inference was based on the features extracted for each $50\,\mathrm{ms}$. Because the model was trained on mixtures each of which contained at most five speakers, it is challenging to obtain diarization results for more than five speakers. Therefore, we produce diarization results for more than five speakers following this procedure: i) decide the maximum number of speakers $K(\leq 5)$ to estimate, ii) decode at most K speaker's diarization results, iii) stop inference if the estimated number of speakers is less than K otherwise continue to the next step, iv) select frames in which all the decoded speakers are inactive and back to i). We varied $K \in \{1, 2, 3, 4, 5\}$ at the first iteration and fixed it to 5 from the second iteration. Finally, the five estimated results are combined using the modified DOVER-Lap to obtain the final results of the EEND-EDA-based system (**System (3)**).

We also used SC-EEND [12] with the replacement of Transformer encoders with Conformer encoders. We trained the model for 200 epochs using simulated mixtures, each of which contains at most four speakers, and then adapted it on the DIHARD III dev set for another 200 epochs (**System (4)**).

As for post-processing for each system, we filtered false-alarm frames and recovered missed frames by assigning the speakers with the highest posterior probabilities using VAD. We used the oracle VAD for Track 1, and the estimated VAD described in Section 1.1 for Track 2.

1.3. EEND as post-processing

The trained EEND-EDA model in Section 1.2 was also used to refine the diarization results of the x-vector-based system in Section 1.1. We used EEND as post-processing (EENDasP) [13] to update the diarization results of the TDNN-based x-vector system (**System (5)**).

1.4. System fusion

We finally applied DOVER-Lap [9] to combine the diarization results of the x-vector-based two systems, EEND-based two systems, and EENDasP system (**System (6)**). The original DOVER-Lap assigns uniformly-divided regions for each speaker if the multiple speakers are weighted equally in the label voting stage, but we found that this leads to the increase of missed speech. Thus, we assigned all the speakers to the region without any division. We also introduced a weighting mechanism to change the importance of each system manually.

1.5. Self-supervised adaptation (SSA)

After the first system fusion, the EEND-EDA model was re-adapted using the estimated results as pseudo labels of the eval set. The inference results of EEND-EDA and x-vector + EENDasP are updated using the new model (**Systems (7) & (8)**) for the final system fusion (**System (9)**).

2. Experimental results

Table 2 shows DERs and JERs on Tracks 1 and 2. Every subsystem significantly outperformed the baseline system [10]. Our best system achieved $11.58\,\%$ and $14.09\,\%$ of DERs on the full and core evaluation set in Track 1, respectively. It also achieved $16.94\,\%$ and $20.01\,\%$ of DERs in Track 2.

3. References

[1] M. Lavechin *et al.*, "End-to-end domain-adversarial voice activity detection," in *INTERSPEECH*, 2020, pp. 3685–3689.

[2] D. Snyder *et al.*, "MUSAN: A music, speech, and noise corpus," arXiv:1510.08484, 2015.

[3] M. Diez *et al.*, "Speaker diarization based on Bayesian HMM with eigenvoice priors," in *Odyssey*, 2018, pp. 102–109.

[4] D. Snyder *et al.*, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP*, 2019, pp. 5796–5800.

[5] S.-H. Gao *et al.*, "Res2Net: A new multi-scale backbone architecture," *IEEE TPAMI*, 2020.

[6] T. Zhou *et al.*, "ResNeXt and Res2Net structures for speaker verification," in *SLT*, 2021.

[7] J. Deng *et al.*, "ArcFace: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4685–4694.

[8] A. Nagrani *et al.*, "VoxCeleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[9] D. Raj *et al.*, "DOVER-Lap: A method for combining overlap-aware diarization outputs," in *SLT*, 2021.

[10] N. Ryant *et al.*, "The third DIHARD diarization challenge," arXiv:2012.01477, 2020.

[11] S. Horiguchi *et al.*, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *INTERSPEECH*, 2020, pp. 269–273.

[12] Y. Fujita *et al.*, "Neural speaker diarization with speaker-wise chain rule," arXiv:2006.01796, 2020.

[13] S. Horiguchi *et al.*, "End-to-end speaker diarization as post-processing," arXiv:2012.10055, 2020.