CIAIC System Descriptions

1st Meng-Zhen Li
CIAIC, School of Marine Science and Technology
Northwestern Polytechnical University
XiAn, China
limengzhen@mail.nwpu.edu.cn

2nd Yi-jun Gong
CIAIC, School of Marine Science and Technology
Northwestern Polytechnical University
XiAn, China
gongyj@mail.nwpu.edu.cn

***Abstract*—We make use of a x-vector extractor model which was trained by Brno University of Technology during the Second DIHARD Diarization Challenge. The clustering methods we used was agglomerative hierarchical clustering (AHC) and variational Bayes based clustering(VBx). Experimental results show that the VBx systems are better than the AHC systems in both develapment set(DEV) and evaluation set(EVAL).**

***Index Terms*—AHC, VBx, x-vector, Dihard2, Dihard3**

I. Introduction

The x-vector extractor model was trained on the VoxCeleb corpora [1], [2] by Brno University of Technology during the Second DIHARD Diarization Challenge[1]. Two of the PLDA model were trained on VoxCeleb data and the Second DIHARD development set [4] respectively. The clustering methods we used was agglomerative hierarchical clustering (AHC) and variational Bayes based clustering(VBx) [5], [6]. Experimental results show that the VBx systems are better than the AHC systems.

II. Notable highlights

VBx diarization use x-vectors as input features which are generated by a Bayesian hidden Markov model.

III. Data resources

As we mentioned in the first section,our x-vector extractor model was trained on data from the VoxCeleb corpora and used the Kaldi toolkit.

VoxCeleb[2] contains over 100,000 utterances for 1,251 celebrities, extracted from videos uploaded to YouTube. The dataset is gender balanced, with 55% of the speakers male. The speakers span a wide range of different ethnicities, accents, professions and ages. There are no overlapping identities between development and test sets.

We also used two probabilistic linear discriminant analysis (PLDA) models, one trained on VoxCeleb data and another on the DIHARD development set.

IV. Detailed description of algorithm

The VBx diarization consists of the following five aspects.

- computing fbank features
- computing x-vectors
- doing Agglomerative Hierarchical Clustering on x-vectors as a first step to produce an initialization
- apply Variational Bayes HMM over x-vectors to produce the diarization output
- score the diarization output

The parameter we use is alpha=0.55, thr=0.0, tareng=0.3, smooth=5.0, lda-dim=220, Fa=0.4, Fb=11, loopP=0.80.

Table I shows the diarization result on the DIHARD 2020 eval set.

TABLE I
DIHARD 2020 EVAL SET

Algorithm	Dataset	DER(%)	JER(%)
AHC	Full	20.16	34.79
	Core	21.77	38.46
VBx	Full	15.67	33.80
	core	16.68	37.91

V. Results on the development set

Table II demonstrates the results on the development set. From the table we can see that JER results are higher than DER. In general, the DER and JER are positively correlated. The performance of VBx clustering are better than AHC in several both full set and core set.

Table III and IV shows the diarization results under different domain conditions.

References

[1] Nagrani, Arsha , J. S. Chung , and A. Zisserman . ”VoxCeleb: a large-scale speaker identification dataset.” Interspeech 2017.

[2] Chung, Joon Son , A. Nagrani , and A. Zisserman . ”VoxCeleb2: Deep Speaker Recognition.” (2018).

TABLE II
DIHARD 2020 DEV SET

Algorithm	Dataset	DER(%)	JER(%)
AHC	Full	21.60	36.98
	Core	22.24	39.94
VBx	Full	16.37	34.40
	Core	16.80	37.99

[1] https://github.com/BUTSpeechFIT/VBx

[2] https://github.com/cyrta/voxceleb

TABLE III

DIHARD 2020 DEV SET(FULL) RESULTS BASED ON VBX DIARIZATION

Metrics	All	Audiobooks	Broadcast	Clinical	Courtroom	CTS	Map	Meeting	Restaurant	field	lab	Web
DER(%)	16.37	2.35	2.64	10.86	2.95	16.21	4.90	26.48	43.68	12.24	8.26	35.05
JER(%)	34.40	2.43	15.92	15.50	13.57	20.92	8.76	46.21	68.96	39.77	11.87	67.79
B3-Precision	0.80	1.00	0.97	0.89	0.95	0.72	0.93	0.60	0.47	0.85	0.90	0.70
B3-Recall	0.91	0.97	0.99	0.92	0.99	0.87	0.96	0.87	0.79	0.94	0.94	0.90
B3-F1	0.85	0.99	0.98	0.91	0.97	0.79	0.95	0.71	0.59	0.89	0.92	0.79
GKT(ref,sys)	0.91	0.97	0.99	0.92	0.99	0.87	0.96	0.86	0.78	0.94	0.94	0.89
GKT(sys, ref)	0.80	1.00	0.97	0.89	0.95	0.72	0.93	0.60	0.47	0.85	0.90	0.70
H(ref|sys)	0.63	0.00	0.11	0.31	0.19	0.74	0.22	1.44	2.06	0.51	0.31	1.05
H(sys|ref)	0.24	0.07	0.05	0.21	0.05	0.34	0.11	0.37	0.58	0.18	0.17	0.26
MI	9.14	4.29	5.06	6.92	5.80	6.91	5.83	5.35	4.96	5.04	5.29	5.70
NMI	0.95	0.99	0.98	0.96	0.98	0.93	0.97	0.86	0.80	0.94	0.96	0.90

TABLE IV

DIHARD 2020 DEV SET(FULL)RESULTS BASED ON AHC DIARIZATION

Metrics	All	Audiobooks	Broadcast	Clinical	Courtroom	CTS	Map	Meeting	Restaurant	field	lab	Web
DER(%)	21.60	34.55	2.68	18.47	3.51	20.08	6.90	28.63	46.20	16.13	12.68	44.09
JER(%)	36.98	33.72	16.12	21.52	14.59	24.85	11.85	47.21	68.77	41.83	17.27	67.73
B3-Precision	0.80	1.00	0.97	0.90	0.95	0.72	0.93	0.60	0.48	0.85	0.89	0.72
B3-Recall	0.84	0.60	0.99	0.85	0.98	0.82	0.94	0.84	0.75	0.89	0.89	0.73
B3-F1	0.82	0.75	0.98	0.87	0.96	0.77	0.93	0.70	0.58	0.87	0.89	0.73
GKT(ref,sys)	0.84	0.59	0.99	0.85	0.98	0.81	0.94	0.83	0.74	0.88	0.88	0.73
GKT(sys, ref)	0.80	1.00	0.97	0.90	0.95	0.72	0.92	0.60	0.47	0.85	0.89	0.72
H(ref|sys)	0.62	0.00	0.12	0.30	0.20	0.74	0.24	1.43	2.00	0.50	0.32	0.91
H(sys|ref)	0.47	1.21	0.04	0.44	0.07	0.53	0.18	0.48	0.74	0.35	0.34	0.80
MI	9.15	4.29	5.05	6.93	5.79	6.91	5.80	5.37	5.02	5.05	5.28	5.84
NMI	0.94	0.88	0.98	0.95	0.98	0.92	0.97	0.85	0.79	0.92	0.94	0.87

[3] Povey, Daniel , et al. ”The Kaldi Speech Recognition Toolkit.” Idiap (2012).

[4] Ryant, Neville , et al. ”The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines.” (2019).

[5] Landini, Federico , et al. ”BUT System for the Second DIHARD Speech Diarization Challenge.” 2020.

[6] Diez, Mireia , et al. ”Optimizing Bayesian Hmm Based X-Vector Clustering for the Second Dihard Speech Diarization Challenge.” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2020.