The Kriston AI System for the VoxCeleb Speaker Recognition Challenge 2022：Track4

Qutang Cai, Guoqiang Hong, Zhijian Ye, Ximin Li, Haizhou Li

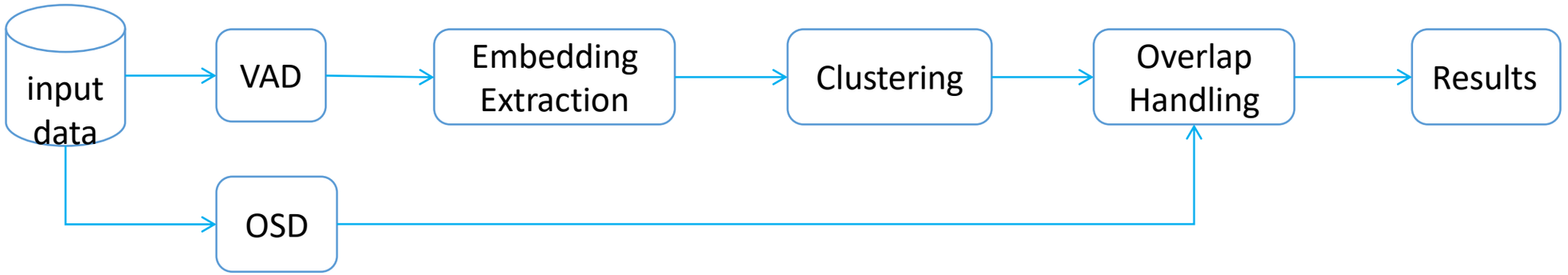
Kriston AI Lab

Presented by **Guoqiang Hong**

September 22, 2022

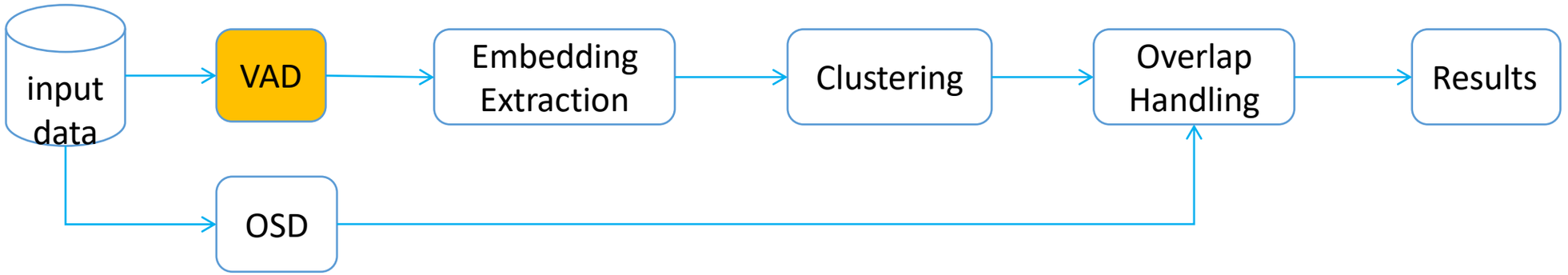
CONTENTS

1.System Overview



1. Voice activity detection(VAD)
2. Speaker embedding extraction
3. Clustering
    3.1 Agglomerative hierarchical clustering(AHC)
    3.2 Variational Bayes hidden Markov model(VB-HMM)
4. Overlap speech detection(OSD)

2.Voice Activity Detection



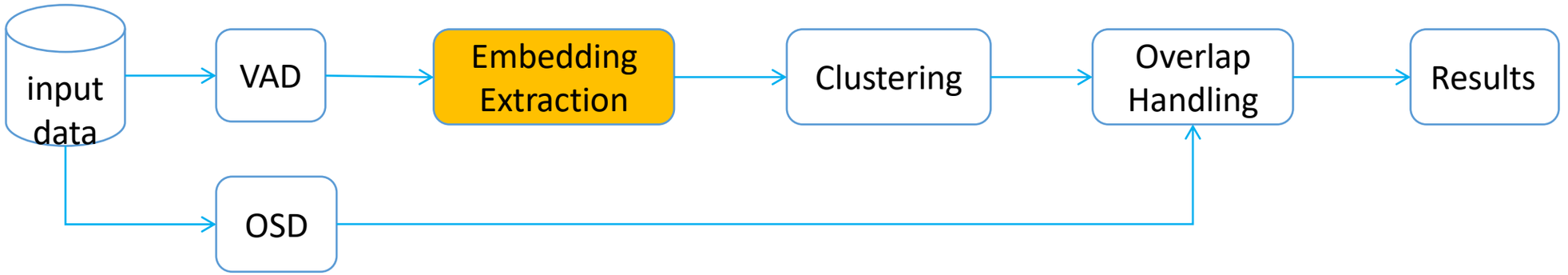
- VAD models like [1] with different acoustic features: 80-dim Fbank, 30-dim MFCC
- Fuse three models with equal weights

Table 1: *The false alarm (FA), miss detection (MISS) and accuracy of the VAD model.*

System	FA[%]	MISS[%]	Accuracy[%]
FBank	3.49	1.49	95.00
MFCC	4.27	0.92	94.80
pyannote	3.22	1.62	95.15
Fusion	3.55	1.06	95.37

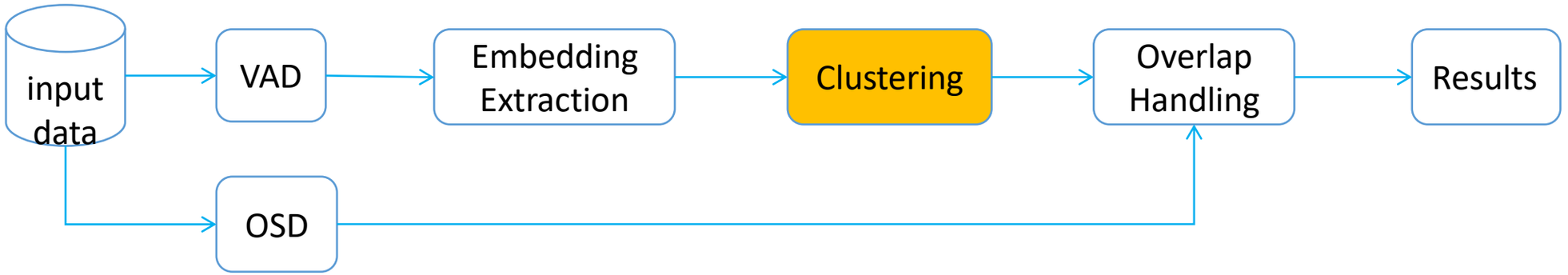
[1] W. Wang, D. Cai, Q. Lin, L. Yang, J. Wang, J. Wang, and M. Li, “The dku-dukeece-lenovo system for the diarization task of the 2021 voxceleb speaker recognition challenge,” 2021.

3.Embedding Extraction



- Model: R6 trained for track 1
- Evaluation: EER=0.44% on VoxCeleb1-O, cosine similarity
- Segment: 1.5s duration, 0.25s step

4.Clustring



Initial-clustering:

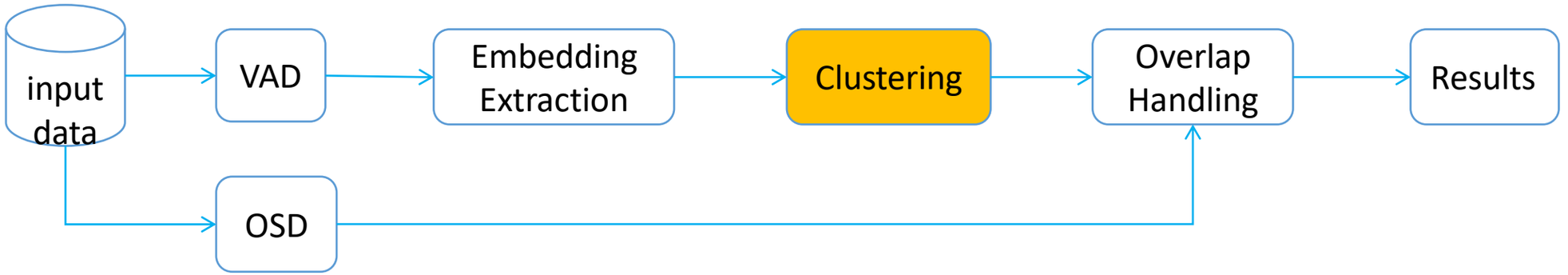
- AHC[2] using cosine similarity

Re-clustering:

- VB-HMM[3] using cosine similarity
- Score calibration using AS-Norm[4]

[2] F. Landini, O. Glembek, P. Matejka, J. Rohdin, L. Burget, M. Diez, and A. Silnova, “Analysis of the but diarization system for voxconverse challenge,” 2020.
[3] F. Landini, J. Profant, M. Diez, and L. Burget, “Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization:theory, implementation and analysis on standard tasks,”2020
[4] P. Matejka, O. Novotny, O. Plchot, L. Burget, M. Diez, and ´J. Cernock ˇ y, “Analysis of score normalization in multilingual ´speaker recognition,” in Proc. Interspeech, 2017, pp. 1567–1571.

4.Clustring(VB-HMM)



Modify equation (16)-(18) in [3]:

$$\alpha_s = \frac{F_A}{F_B} L_s^{-1} \sum_t \gamma_{ts} \rho_t$$

$$\alpha_s = \frac{F_A}{F_B} L_s^{-1} \sum_t \gamma_{ts} \rho_t$$

$$L_s = \mathrm{I} + \frac{F_A}{F_B} \left(\sum_t \gamma_{ts} \right) \Phi$$

$$L_s = \mathrm{I} + \frac{F_A}{F_B} \sum_t \gamma_{ts}$$

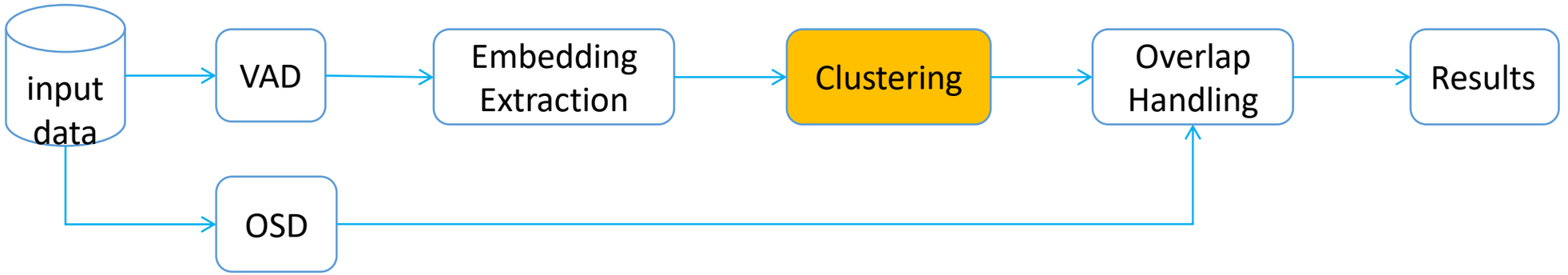
$$\rho_t = \mathrm{V}^{\mathrm{T}} x_t$$

$$\rho_t = x_t = F_C E_t$$

where E_t is the L2-normalized speaker embedding at frame t, F_C is a scale parameter

[3] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization:theory, implementation and analysis on standard tasks,"2020

4.Clustring(VB-HMM with As-Norm)



Replace the $\alpha_s^T \rho_t$ and Φ terms in $\log\bar{p}(x_t \mid s)$ (equation (23) in [3]):

$$\alpha_s^T \rho_t = \frac{F_A F_C^2}{F_B} l_s^{-1} \frac{\beta_s^T E_t - \mu_s}{\delta_s} \sum_t \gamma_{ts}$$

$$\Phi = \mathrm{I}$$

$$\beta s = \frac{\sum_t \gamma_{ts} E_t}{\sum_t \gamma_{ts}}$$

$$l_s = 1.0 + \frac{F_A}{F_B} \sum_t \gamma_{ts}$$

where μ_s and σ_s are mean and standard deviation of β_s

[3] F. Landini, J. Profant, M. Diez, and L. Burget, “Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization:theory, implementation and analysis on standard tasks,”2020

4.Clustring

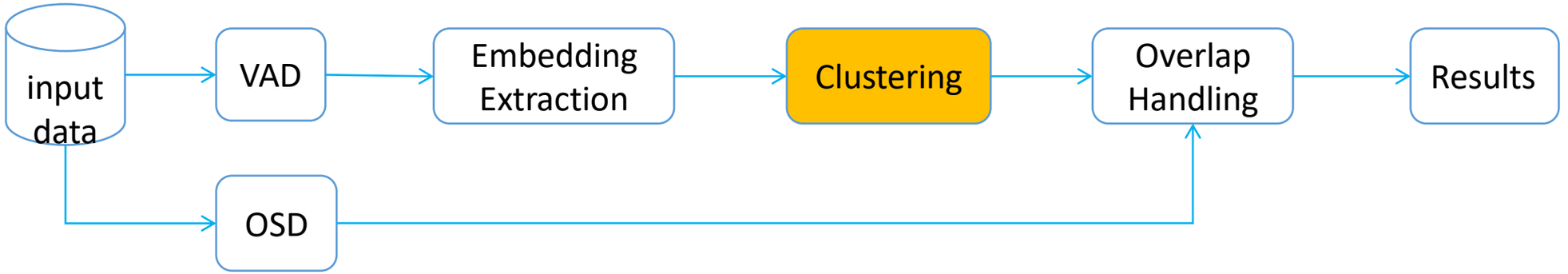
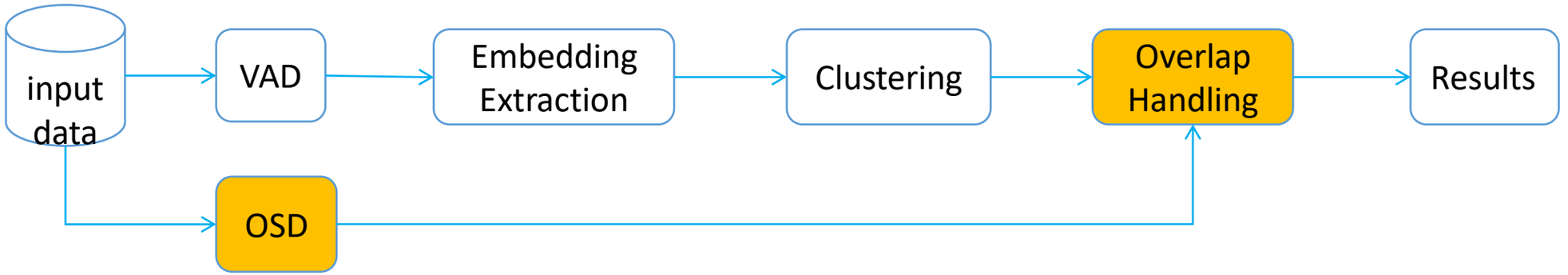


Table 2: *The DER and JER of the proposed speaker diarization system on the test set of VoxConverse.*

System	DER[%]	JER[%]
VB	4.42	26.43
VB+asnorm	4.29	26.81

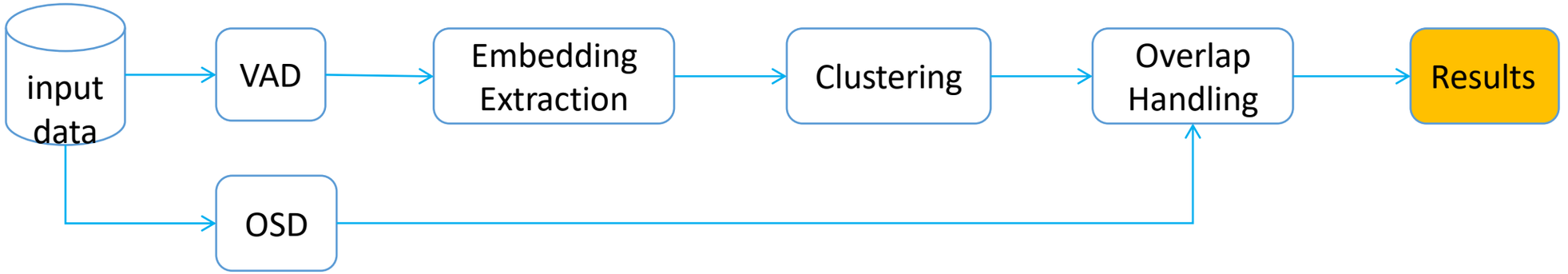
5.Overlap And Handling



- OSD: similar to VAD
- Handling: find the two closest speakers in time[5]

[5] Wang, Keke and Mao, Xudong and Wu, Hao and Ding, Chen and Shang, Chuxiang and Xia, Rui and Wang, Yuxuan, “The ByteDance Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021,” 2021

6.Results



- Our best system obtained 4.86% DER and 25.48% JER

Reference

[1] W. Wang, D. Cai, Q. Lin, L. Yang, J. Wang, J. Wang, and M. Li, “The dku-dukeece-lenovo system for the diarization task of the 2021 voxceleb speaker recognition challenge,” 2021. [Online]. Available: https://arxiv.org/abs/2109.02002

[2] F. Landini, O. Glembek, P. Matejka, J. Rohdin, L. Burget, M. Diez, and A. Silnova, “Analysis of the but diarization system for voxconverse challenge,” 2020. [Online]. Available: https://arxiv.org/abs/2010.11718

[3] F. Landini, J. Profant, M. Diez, and L. Burget, “Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization:theory, implementation and analysis on standard tasks,”2020.[Online]. Available: https://arxiv.org/abs/2012.14952

[4] P. Matejka, O. Novotny, O. Plchot, L. Burget, M. Diez, and ´J. Cernock ˇ y, “Analysis of score normalization in multilingual ´speaker recognition,” in Proc. Interspeech, 2017, pp. 1567–1571.

[5] Wang, Keke and Mao, Xudong and Wu, Hao and Ding, Chen and Shang, Chuxiang and Xia, Rui and Wang, Yuxuan, “The ByteDance Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021,” 2021

THANK YOU