大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

icassp 2022 Singapore

Attention Back-end for Automatic Speaker Verification with Multiple Enrollment Utterances

Chang Zeng[1,2], Xin Wang[1], Erica Cooper[1], Xiaoxiao Miao[1], Junichi Yamagishi[1,2]

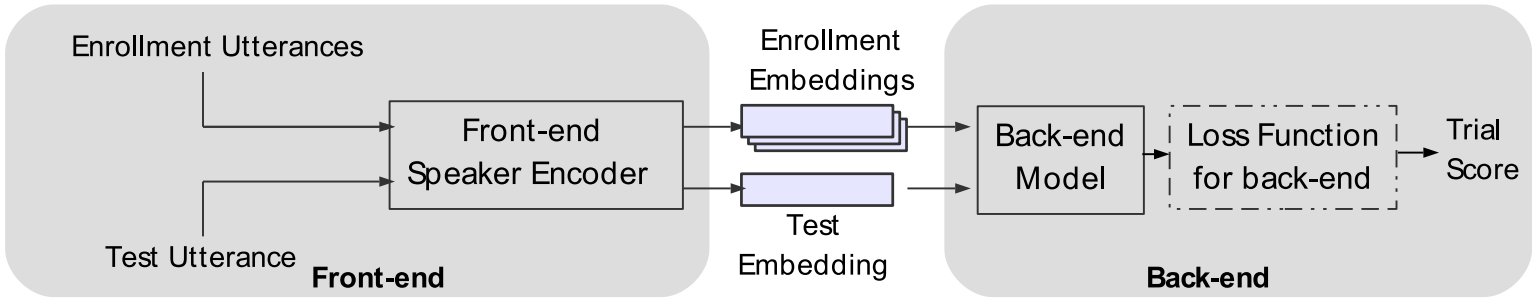
[1]National Institute of Informatics, Japan [2]SOKENDAI, Japan

Monday, 9th May, 2022

Outline

1. Introduction

- ASV framework
  - Front-end speaker encoder
    - TDNN, ECAPA-TDNN
    - ResNet series
  - Back-end model
    - Cosine similarity
    - Probabilistic linear discriminant analysis (PLDA)
    - Neural PLDA (NPLDA) [1]



- Single enrollment VS Multiple enrollments
  - The case of multiple enrollments is more robust in real world setting;
  - Multiple enrollments contain more acoustic variations of enrolled speaker.
- Common operation for Multiple enrollments case
  - Concatenate waveforms or acoustic features
  - Average speaker embeddings

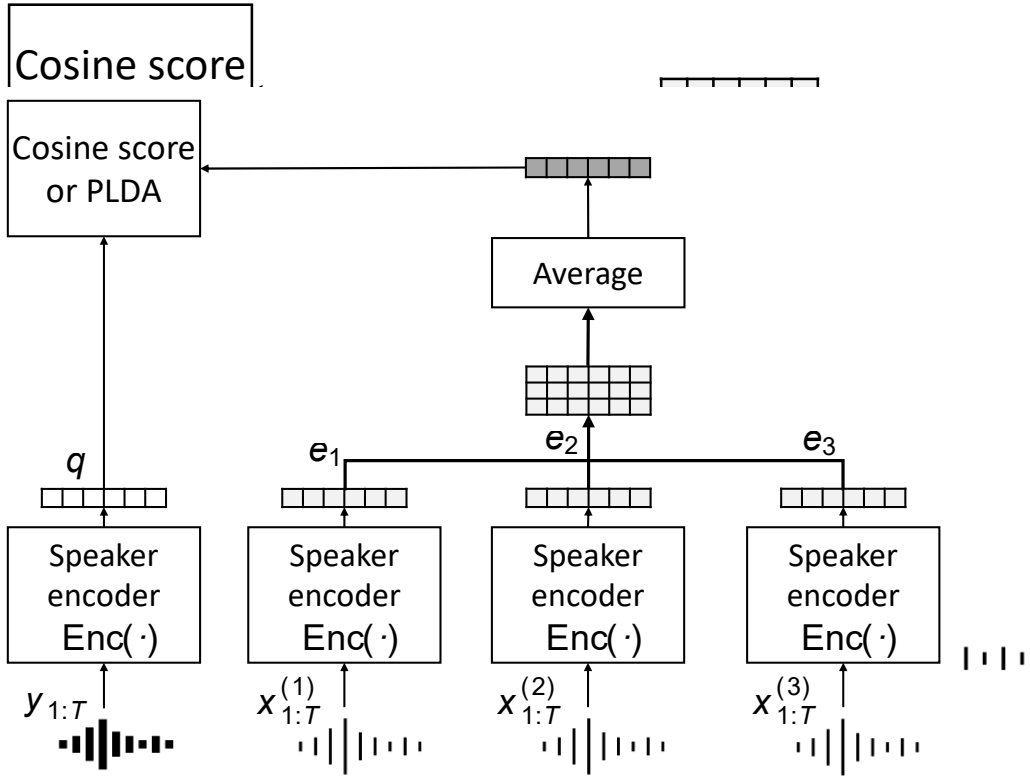
1. Introduction

- Can we use multiple enrollments more efficiently in a neural back-end?
  - Learning intra-relationships among speaker embeddings of multiple enrollments;
  - Aggregating a varying number of enrollment speaker embeddings by adaptive weights.

Outline

2. Conventional Back-end Models

- Multiple enrollments case, where enrollment speaker has an enrollment set including K utterances. Note K is a varying number.
  - Concatenating all enrollment utterances as a long waveforms
  - Extracting speaker embedding from each utterance, then averaging all speaker embeddings

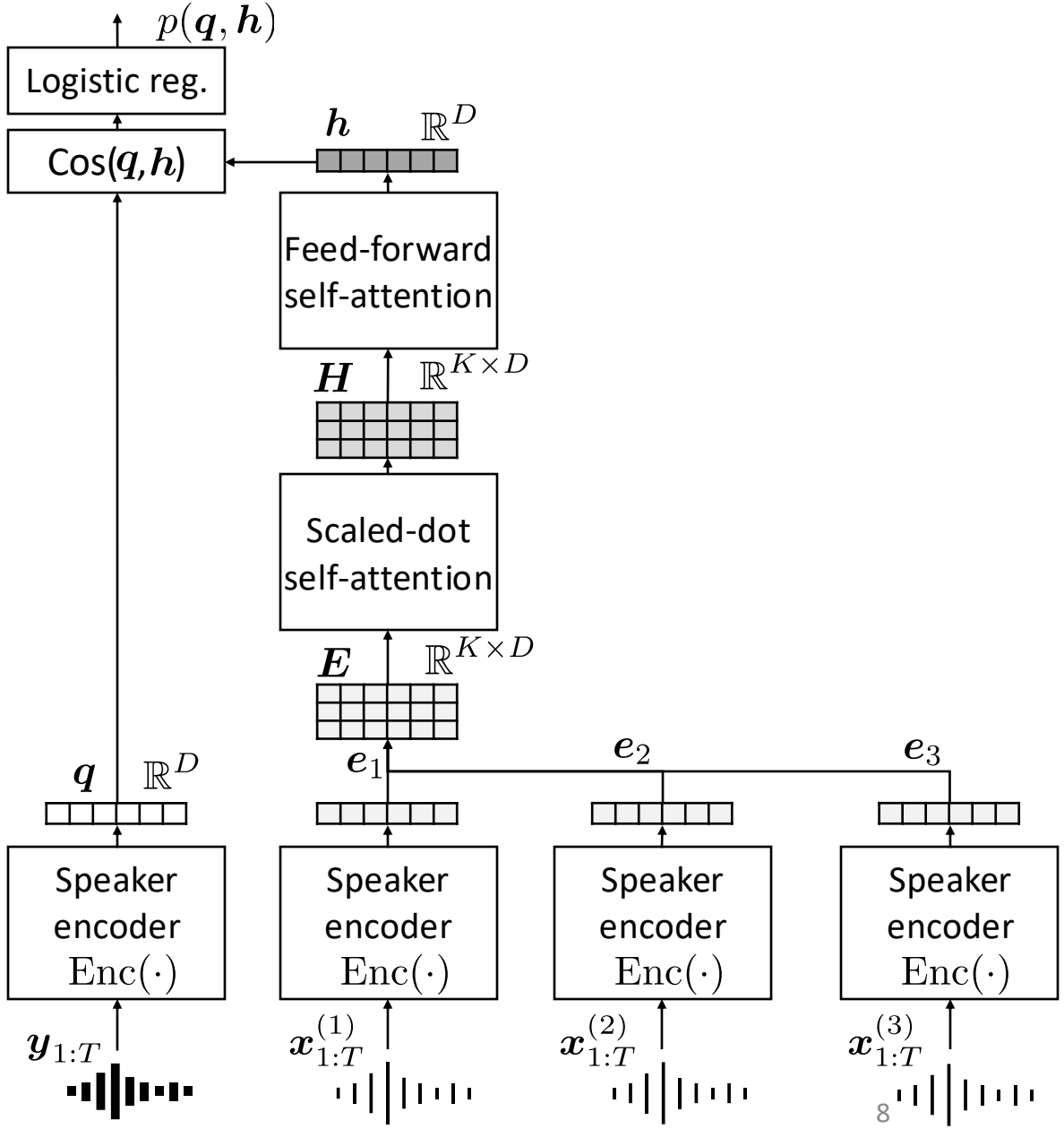


Note: The scoring method of PLDA can be extended to adapt the case of multiple enrollments in theoritically except concatenation or average operation. As for the detail, please refer to [2,3]

Outline

3. Attention Back-end Model

- Model architecture
  - Extract enrollment speaker embeddings and testing speaker embedding $\boldsymbol{q}$
  - Stacking all enrollment speaker embeddings as matrix $\boldsymbol{E}$.
  - Exploring intra-relationships among all enrollment speaker embeddings
    - Scaled-dot self-attention (SDSA) [3]
  - Aggregating a varying number of enrollment speaker embeddings by adaptive weights
    - Feed-forward self-attention (FFSA) [4]
  - Score calibration
    - Logistic regression (LR) for score calibration



3. Attention Back-end Model

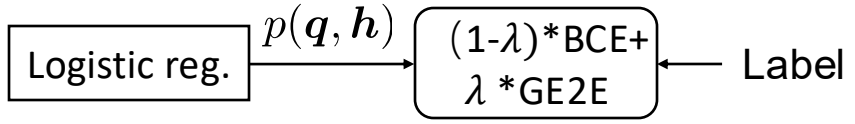
- Trials sampling method for training
  - Why: Introduce multiple enrollment process in training stage
  - How:
    - Load speaker-balanced min-batch from dataset
    - Rearrange the mini-batch to form trial pairs (*testing, enrollments*)
    - Positive pair: test trial and enrollment data are from the same speaker, one test trial is selected from the speaker's data, and the rest are left for enrollment.
    - Negative pair: pairs marked by test trial= ✓, enroll=(×, ×, ×) of other speakers included in a mini-batch.

Table 1. Composition of pairs of (test-trial, enrollment-data) for training back-end model and ground-truth labels from mini-batch. A, B, and C are speaker IDs, and 1, 2, 3 and 4 are his or her audio IDs. ✓and ×denote test and enrollment audio files, respectively.

	A				*B*				*C*						
	1	2	3	4	1	2	3	4	1	2	3	4	Test	Enroll	Label
Trials to be used for training							⋮								

3. Attention Back-end Model

- Loss functions
  - A weighted sum of binary cross-entropy (BCE) loss and generalized end-to-end (GE2E) loss
  - Binary cross-entropy loss
  - Generalized end-to-end loss



Trials to be used for training													Test	Enroll	Label
	A				*B*				*C*						
	1	2	3	4	1	2	3	4	1	2	3	4			
	✓	×	×	×									$\boldsymbol{q}_{A1}$	$\boldsymbol{h}_{A1}$	P
	✓					×	×	×					$\boldsymbol{q}_{A1}$	$\boldsymbol{h}_{B1}$	N
	✓									×	×	×	$\boldsymbol{q}_{A1}$	$\boldsymbol{h}_{C1}$	N
	×	✓	×	×									$\boldsymbol{q}_{A2}$	$\boldsymbol{h}_{A2}$	P
		✓			×		×	×					$\boldsymbol{q}_{A2}$	$\boldsymbol{h}_{B2}$	N
		✓							×		×	×	$\boldsymbol{q}_{A2}$	$\boldsymbol{h}_{C2}$	N
							⋮								
	×	×	×									✓	$\boldsymbol{q}_{C4}$	$\boldsymbol{h}_{A4}$	N
					×	×	×					✓	$\boldsymbol{q}_{C4}$	$\boldsymbol{h}_{B4}$	N
									×	×	×	✓	$\boldsymbol{q}_{C4}$	$\boldsymbol{h}_{C4}$	P

Outline

- 1. Introduction
- 2. Conventional Back-end Models
- 3. Attention Back-end Model
  - 3.1 Model architecture
  - 3.2 Trials sampling method
  - 3.3 Loss functions
- **4. Experimental Results**
  - 4.1 Datasets
  - 4.2 Result and analysis
  - 4.3 Ablation study
- 5. Conclusions

4. Experimental Results

- **Datasets**

  - Single enrollment case

    VoxCeleb1&2 [5,6]

7,000 +

speakers

VoxCeleb contains speech from speakers spanning a wide range of different ethnicities, accents, professions and ages.

1 million +

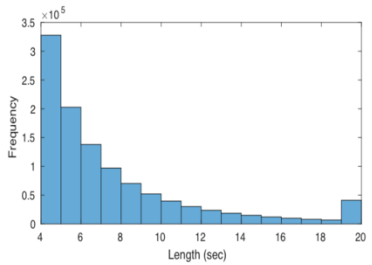
utterances

All speaking face-tracks are captured "in the wild", with background chatter, laughter, overlapping speech, pose variation and different lighting conditions.

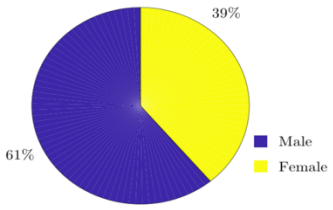
2,000 +

hours

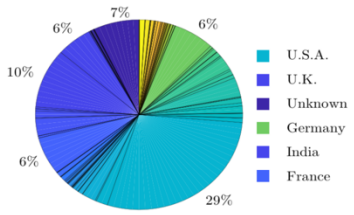
VoxCeleb consists of both audio and video. Each segment is at least 3 seconds long.



Utterance Lengths



Gender Distribution



Nationality Distribution

source:https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

Note: The number of enrollment utterances of the evaluation protocol in CNCeleb dataset is varying!

- Multiple enrollments case

  CNCeleb1&2 [7,8]

3,000

Speakers

CN-Celeb contains speech from Chinese celebrities.

600,000 +

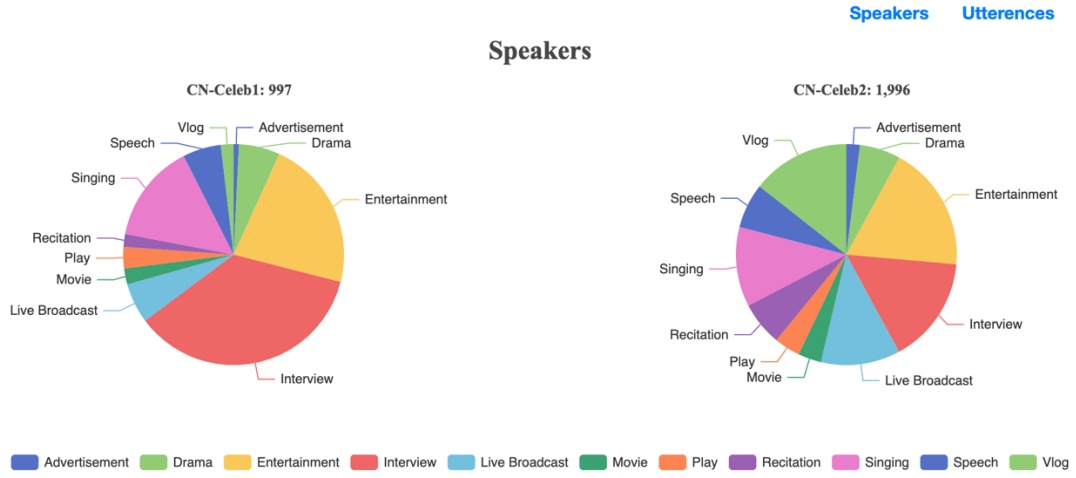
Utterances

CN-Celeb covers multiple genres of speech, including entertainment, interview, singing, play, movie, vlog, live broadcast, speech, drama, recitation advertisement.

1,200 +

Hours

CN-Celeb consists of complex long-short challenge which meets the scenarios of most



source:http://cnceleb.org/dataset

4. Experimental Results

- Speaker encoders
  - TDNN with statistics pooling (TDNN)
  - ResNet34

  } Speaker encoder without frame-level attention

  - TDNN with attentive statistics pooling (TDNN-ASP)
  - ECAPA-TDNN

  } Speaker encoder with frame-level attention

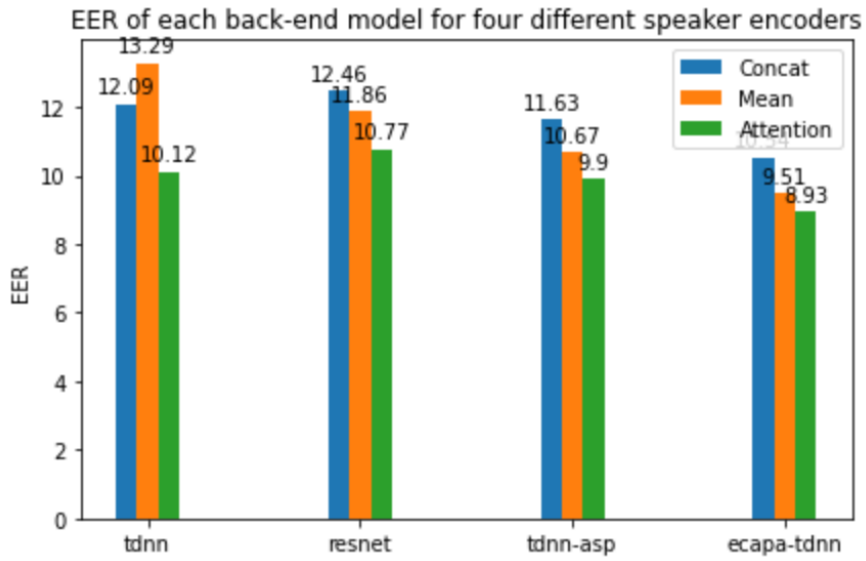
4. Experimental Results

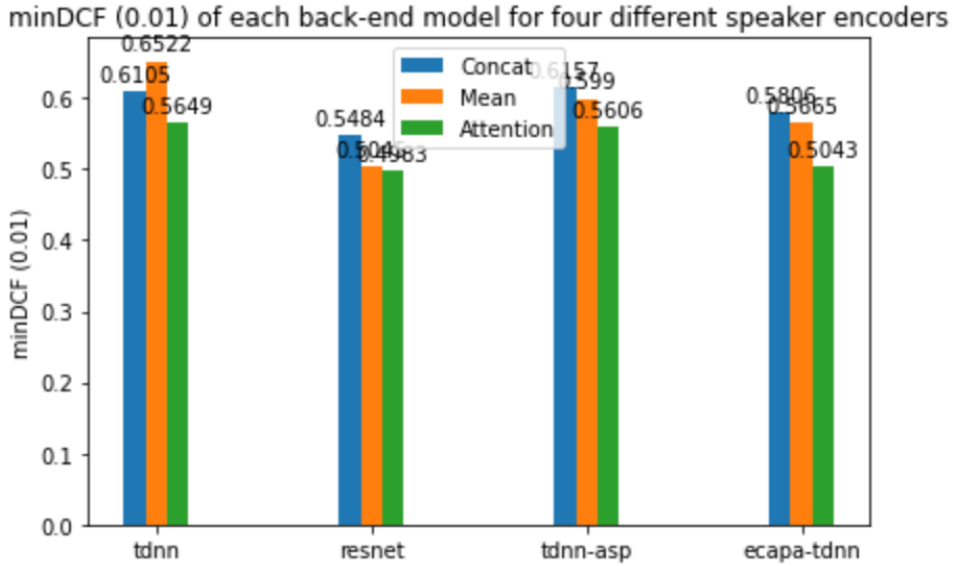
- Result and analysis
  - Single enrollment case on VoxCeleb
    - Our back-end model is comparable with PLDA in term of EER, even if there is no intra-relationship in enrollment utterance;
    - As for minDCF, the proposed model slightly outperforms PLDA, which proves that our model works reasonably well for the single enrollment case.

Table 2. VoxCeleb verification performance results with TDNN and different back-ends

Back-end	EER(%)	minDCF(0.01)	minDCF(0.001)
Cosine	10.51	0.7928	0.8718
PLDA	**3.14**	0.3456	0.5567
Proposed	3.26	**0.3323**	**0.5134**

  - Multiple enrollments case on CNCeleb
    - For Concat or Mean operation, we give the best performance of cosine similarity or PLDA model.
    - Despite of the criterion, our proposed attention back-end realizes the best performance.





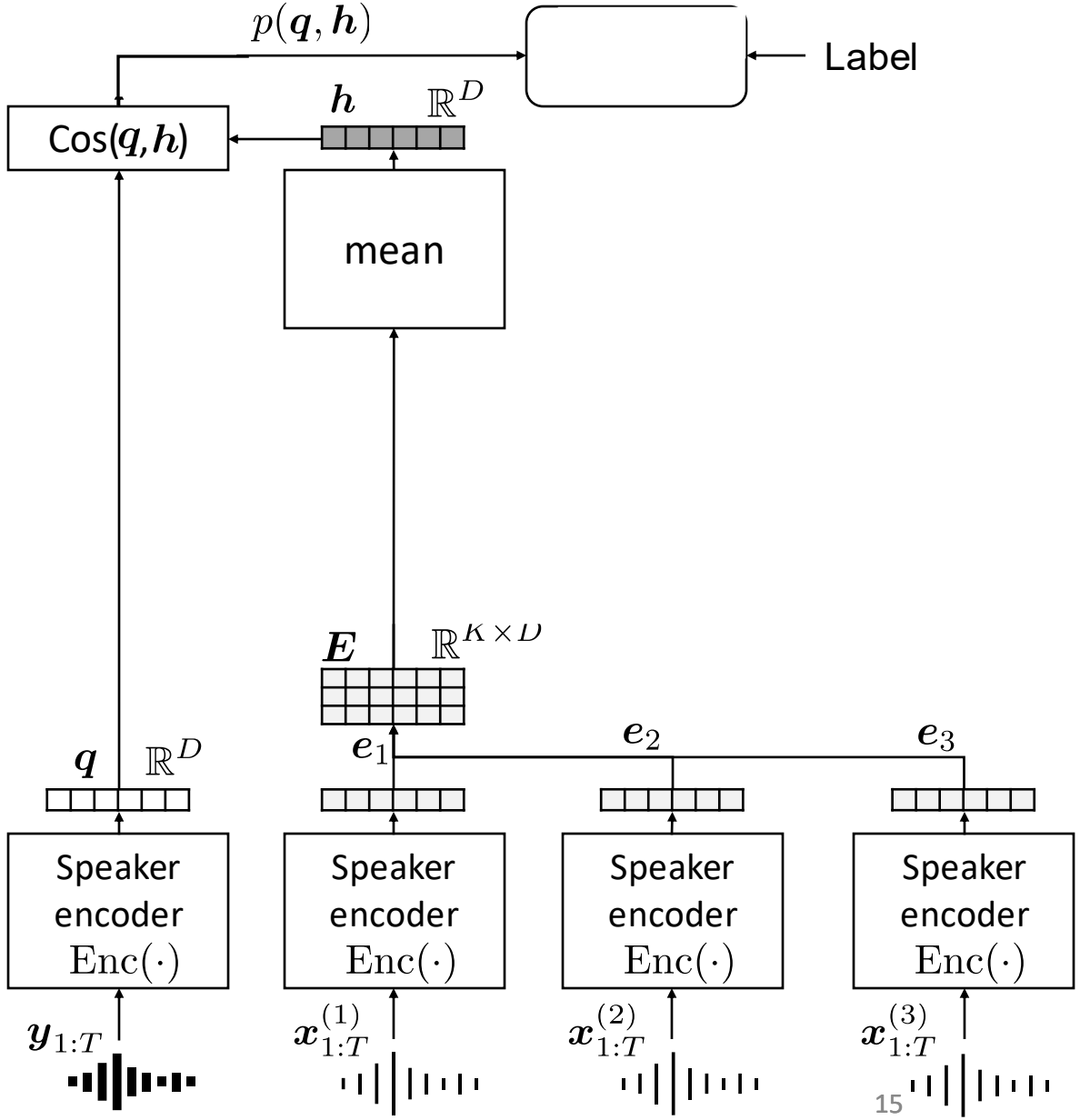
4. Experimental Results

- Ablation study

Table 4. Ablation experiment results on CNCeleb with TDNN speaker encoder.

Systems	EER(%)	DCF(0.01)	DCF(0.001)
SDSA + FFSA + LR + BCE + GE2E	10.12	0.5649	0.7127
FFSA + LR + BCE + GE2E	19.55	0.8095	0.9046
SDSA + mean + LR + BCE + GE2E	10.26	0.5876	0.7270
SDSA + FFSA + BCE + GE2E	48.32	1.0000	1.0000
SDSA + FFSA + LR + GE2E	12.74	0.5937	0.7350
SDSA + FFSA + LR + BCE	10.28	0.5896	0.7402

- From the second and the third lines of the table, two self-attention mechanisms both contributed to the improvement.
- From the fourth line, we can also see that LR for calibration was another essential component of the proposed model
- From a comparison of the fifth and sixth lines, we see that the proposed model using BCE loss only had a lower EER than that using GE2E loss only, while one using GE2E loss resulted in a lower minDCF(0.001) value. Combining BCE and GE2E thus resulted in the lowest values for all the metrics.



Outline

5. Conclusion

- Conventional back-end model like cosine similarity and PLDA cannot work well in multiple enrollments scenario.
- Utterance-level attention mechanism can improve the performance further, even if frame-level attention mechanism and spatial attention mechanism have been applied in speaker encoder.

6. References

1) S. Ramoji, P. Krishnan, and S. Ganapathy, “Neural plda modeling for end-to-end speaker verification,” arXiv preprint arXiv:2008.04527, 2020.

2) Kong Aik Lee, Anthony Larcher, Chang Huai You, Bin Ma, and Haizhou Li, “Multi-session plda scoring of i-vector for partially open-set speaker detection,” in Interspeech, 2013.

3) Padmanabhan Rajan, Anton Afanasyev, Ville Hautam¨aki, and Tomi Kinnunen, “From single to multiple enrollment i-vectors: Practical plda scoring variants for speaker verification,” Digital Signal Processing, vol. 31, pp. 93–101, 2014.

4) Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in NIPS, 2017.

5) Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio, “A structured self-attentive sentence embedding,” in ICLR, 2017.

6) A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in Interspeech, 2017.

7) J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in Interspeech, 2018.

8) Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang, “Cn-celeb: a challenging chinese speaker recognition dataset,” in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7604–7608.

9) Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, “Cn-celeb: multi-genre speaker recognition,” arXiv preprint arXiv:2012.12468, 2020.

Thanks!