The Microsoft System for VoxCeleb Speaker Recognition Challenge 2022

Gang Liu, Tianyan Zhou, Yong Zhao,

Yu Wu, Zhuo Chen, Yao Qian, Jian Wu

Microsoft

09-22-2022

Outline

- Supervised models
- Self-supervised learning (SSL)-based models
- Score calibration and Fusion
- Experiment Results
- Conclusion

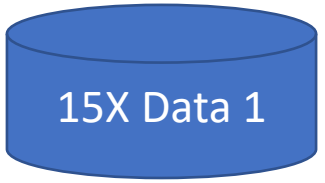
Supervised Models

- Backbone networks:
  - Res2Net-50, Res2Net-101, Res2Net-152
  - ECAPA-TDNN (C=512), ECAPA-TDNN (C=1024)

- Two-stage training strategy
  - Stage 1: 3x speed perturbed Speaker augmentation, 17,982 Speakers, 2s crop length, margin 0.2 or 0.3 for AM and AAM loss, respectively.
  - Stage 2: w/o speed perturbed data, 5,994 Speakers, 6s crop length, margin 0.4 or 0.5 for AM and AAM loss, respectively.

Training Data: VoxCeleb2 Dev

- 2x Speed Augmentation: 10% Speed-up/down
- Kaldi-style data simulation with 4 styles:
  - babble, music, noise, and reverberation.
- On-line data simulation



Self-Supervised Learning (SSL) - Based Models

SSL- based models:

- SSL- Pretrained models:
  - Wav2vec
  - wavLM
- ECAPA-TDNN

- Stage 1: 10 epochs, AAM loss m=0.2, chunk size=3s (the upper stream SSL model are fixed)
- Stage 2: 5 epochs, AAM loss m=0.2, chunk size=3s
- Stage 3: 2~3 epochs, AAM loss m=0.4~0.5, chunk size=6s

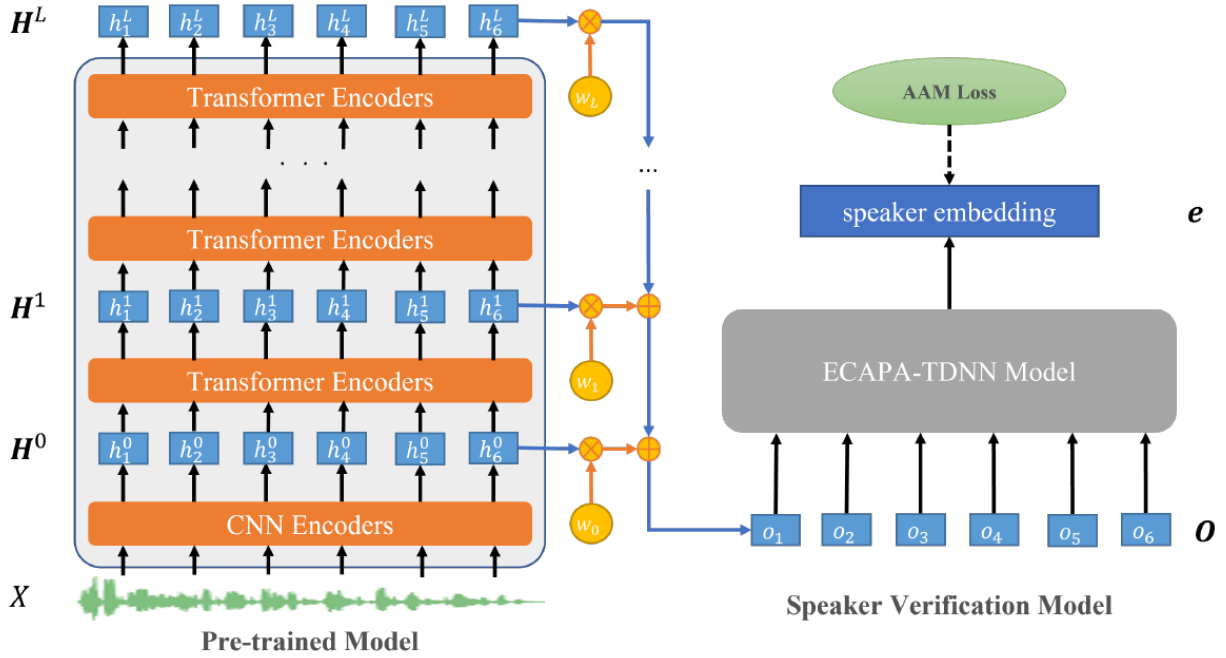


Figure from [1]

[1] "Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?" https://arxiv.org/pdf/2204.12765.pdf

Pretraining Dataset

- SSL-model pre-training data:
  - wav2vec [1]
  - wavLM [2]

VoxCeleb2 Dev: 2.4khrs

model	Data Duration(hrs)	Data Source
Wav2Vec2.0 Large (XLSR)	56k	Multilingual LibriSpeech, CommonVoice, BABLE, Over 36 languages
wavLM-large-v1	94k	LibriLight,VoxPopuli, and GigaSpeech

- Validation data:
  - Vox1-O
  - Vox2-test
  - VoxSRC-22-val
  - VoxSRC-21-val

dataset	#utt	#trial	#target	#nontarget	#speaker
vox1-O	4708	37611	18802	18809	40
vox2-test	26591	30000	15190	14810	120
voxSRC21_val	64711	60000	29969	30031	1251
voxSRC22_val	110366	306432	159789	146643	*1205

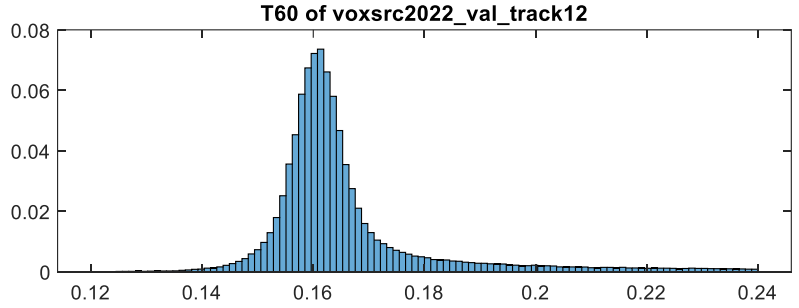
Note: *1205 only counts the speaker from VoxCeleb1 Dev

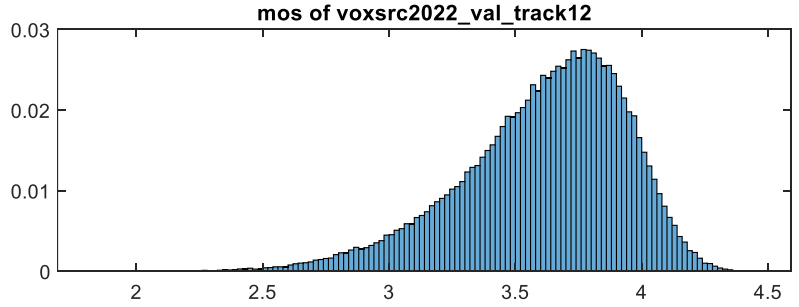
[1] https://dl.fbaipublicfiles.com/fairseq/wav2vec/xlsr_53_56k.pt
[2] https://arxiv.org/pdf/2204.12765.pdf

Score Calibration and Fusion

- SNORM
- Calibration
  - Quality measures: duration, embedding magnitude, cohort mean
- Score fusion
  - Ensemble 13 supervised and SSL-based models
  - Quality measures: T60, MOS

noise





age

Ablation Study on Single System (wavLM-large-v5)

score-processing	Vox1-O		vox2-test		voxSRC-21-val		voxSRC-22-val	
	EER(%)	dcf	EER(%)	dcf	EER(%)	dcf	EER(%)	dcf
CDS	0.43	0.028	2.12	0.055	1.71	0.099	1.20	0.083
+SNORM	0.38	0.024	2.11	0.052	1.53	0.084	1.09	0.073
++Calibration	0.28	0.021	2.10	0.048	1.44	0.080	0.92	0.063

Experiment results

Evaluation results on VoxSRC21-Val and VoxSRC22-val test set.
System 1~7 are SSL-based models. System 8~13: supervised models

No.	System	VoxSRC22-val		VoxSRC21-val	
		EER(%)	DCF	EER(%)	DCF
1	wavLM-large-v1	1.44	0.0944	1.92	0.1205
2	wavLM-large-v2	1.39	0.0902	1.88	0.1116
3	wavLM-large-v3	0.95	0.0658	1.31	0.0721
4	wavLM-large-v4	1.11	0.0744	1.52	0.0879
5	wavLM-large-v5	0.92	0.0630	1.37	0.0797
6	wav2vec2.0-XLSR2-v1	1.25	0.0816	1.67	0.1027
7	wav2vec2.0-XLSR2-v2	1.25	0.0847	1.59	0.0971
8	Res2Net-50	1.36	0.0863	1.56	0.0891
9	Res2Net-101-v1	1.24	0.0857	1.65	0.0984
10	Res2Net-101-v2	1.23	0.0834	1.63	0.0986
11	Res2Net-152	1.25	0.0866	1.74	0.0977
12	ECAPA-TDNN(C512)	1.88	0.1305	2.54	0.1509
13	ECAPA-TDNN(C1024)	1.65	0.1039	2.15	0.1302
14	fusion of 13 models	0.81	0.0518	1.12	0.0585
15	fusion of 13 models + 3 audio quality measures	0.79	0.0513	1.10	0.0565

SSL-based Models

Supervised Models

Experiment results

Evaluation results of two submissions on VoxSRC22-Val and VoxSRC22-blind sets.

Team Name: Strasbourg-Spk

No.	System	VoxSRC22-val		VoxSRC-22-blind	
		EER(%)	DCF	EER(%)	DCF
1	#1	0.950	0.0658	1.606	0.0921
10	#15	0.790	0.0513	1.436	0.0728

22%

21%

Conclusion

- The SSL-based models produce superior performance over the conventional supervised models, attributed to learning from large amount of unlabeled data.
- The ensemble system leverages complementary information from multiple supervised and SSL models, further boosting the performance by +20% relative over the single SSL-based model.

Microsoft

Thank You

Q&A