ID R&D System Description to VoxCeleb Speaker Recognition Challenge 2022

Rostislav Makarov, Nikita Torgashov, Alexander Alenin, Ivan Yakovlev, Anton Okhotnikov

September 22, 2022

ID R&D Inc., USA, New York
{makarov,torgashov,alenin,yakovlev,ohotnikov}@idrnd.net

Overview

Introduction

We used a fusion of **deep ResNets** and **Self-Supervised Learning** models trained on a mixture of private large dataset and publicly available VoxCeleb2 for **Track 2**, and a fusion of **ResNets** trained on VoxCeleb2 alone for **Track 1**.

The final submissions achieved the **first** places on the VoxSRC-22 leaderboard for both **Track 1** and **Track 2** with a $minDCF_{0.05}$ of **0.088** and **0.062** respectively.

Datasets

- **Closed & Open Tracks**
  - VoxCeleb1
  - VoxCeleb2
  - VoxSRC22
- **Open Track only**
  - Self-VoxCeleb dataset
- **Augmentation**
  - MUSAN
  - Real RIRs

VoxCelebs

We used **VoxCeleb2-dev** dataset for training the models for **Track 1**.

For **open** condition (Track 2) we used two datasets:
Voxceleb2-dev and **Self-Voxceleb**.

For validation, **VoxCeleb1-test** set and **VoxSRC22 validation** sets were used.

Self-VoxCeleb dataset

Inspired by the idea of the VoxCeleb2 dataset collection, we adopted and modified the collection method to obtain a similar dataset of increased volume, to which we refer as a **Self-VoxCeleb**.

The dataset size overcomes VoxCeleb2 dataset size by a **multiple factor**, and all the videos are licensed under the **CC BY 4.0**.

We did not use any face recognition model and utilized a speech-based filtering only using pre-trained SV model embeddings.

Architectures:

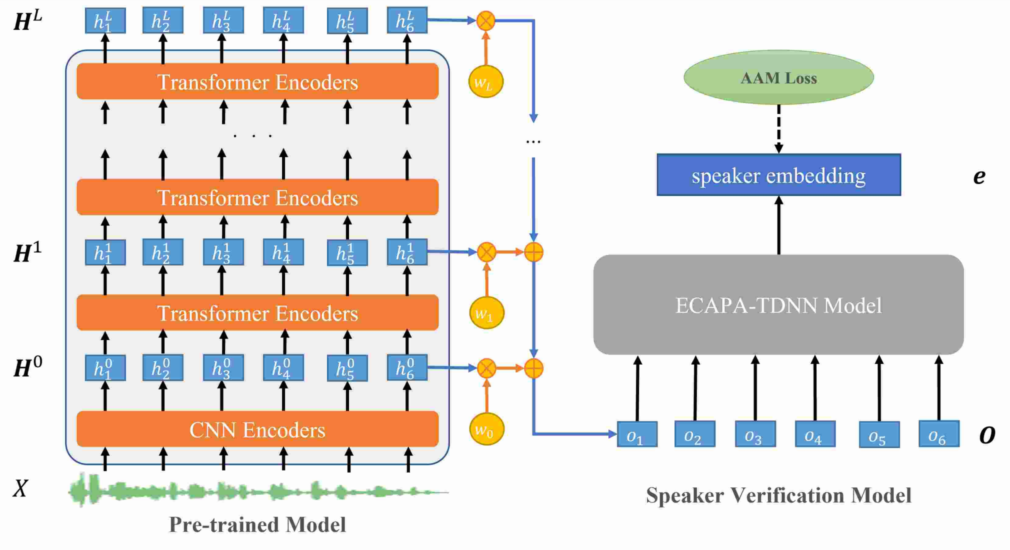
- ResNet
- SSL + ECAPA-TDNN.

As a main architecture we have chosen **ResNet**, that is widely used in speaker recognition and **ECAPA-TDNN** trained on top of the features of self-supervised models, such as **WavLM** and **HuBERT**.

ResNet-202 architecture

Layer name	Structure	Output (C × F × T)
Conv2D	3×3, 128, stride=1	128 × 64 × T
ResBlock-1	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \\ \text{fwSE}, [128, 64] \end{bmatrix} \times 6$	128 × 64 × T
ResBlock-2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \\ \text{fwSE}, [128, 32] \end{bmatrix} \times 16$	128 × 32 × T/2
ResBlock-3	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \\ \text{fwSE}, [128, 16] \end{bmatrix} \times 75$	256 × 16 × T/4
ResBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \\ \text{fwSE}, [128, 8] \end{bmatrix} \times 3$	256 × 8 × T/8
Flatten (C, F)	—	2560 × T/8
StatsPooling	—	5120
Dense	—	256
AM-Softmax	—	Num. of speakers

SSL architecture



Training

Two training stages (ResNet-202 example):

- **Pre-training**
  - Self-VoxCeleb
  - 200 epochs
  - 2 weeks on TPU v3-8 accelerator
- **Fine-tuning**
  - VoxCeleb2 & Self-VoxCeleb
  - No augmentations

Pairwise Scoring & AS-Norm

- For inference, we sliced input samples (both enrollment and verification) into 10×4 seconds chunks resulting in **100** scores as shown in eqs. (1) to (3).

$$N = 10 \tag{1}$$

$$N_{scores} = N \cdot N = 10 \cdot 10 = 100 \tag{2}$$

$$score = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} cosine(enroll_i, verify_j)}{N_{scores}} \tag{3}$$

- The AS-Norm cohort included all **VoxCeleb2-dev** speakers (mean embeddings) with a **top N = 300** trials used to estimate mean and std of scores distribution for normalization.

Quality Measurement Functions: Length

Speech and total length based QMF values were extracted with a help of a standard energy-based Voice Activity Detection (VAD) module. After applying the VAD, we summed all the speech segments lengths into one value.

List of generated and used in out submissions **QMFs**:

- $a)$ speech length of the enrollment model file,
- $b)$ speech length of the trial file,
- $c)$ logarithm of sum of enrollment and trial files speech lengths,
- $d)$ logarithm of sum of enrollment and trial files total lengths.

Quality Measurement Functions: SNR

Signal-to-Noise ratio based QMF values were obtained using the same VAD module. After classifying the voiced and non-voiced segments of a signal, signal-to-noise ratio was calculated using the following eq. (4):

$$SNR_{dB} = 10 \cdot log_{10} \frac{P_{voice}}{P_{non-voice}} \tag{4}$$

where P_{voice} and $P_{non-voice}$ are powers of voiced and non-voiced segments.

We used the following SNR values as **QMF**:

- $e)$ SNR of enrollment model file,
- $f)$ SNR of a trial file.

Quality Measurement Functions: NISQA

NISQA Mean Opinion Score (MOS) was also used in **Track 2** as a QMF term.
It is an open-source model for non-intrusive speech quality estimation.

NISQA predicts the human perception of a speech signal quality on a scale from **1** to **5**.

We utilized the NISQA output for the two following **QMF** values:

- $g)$ NISQA MOS value of enrollment model file,
- $h)$ NISQA MOS value of trial file.

Fusion scheme

The output of our system is an implementation of a linear fusion of cosine similarity scores for all the models and QMF values. To find the weights of each model in a **score-level** fusion we used the **COBYLA** optimizer on **VoxSRC22-dev** set.

The trial score was obtained according to eq. (5):

$$S' = \begin{bmatrix} w_1 \; w_2 \; ... \; w_n \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ ... \\ S_n \end{bmatrix} + \begin{bmatrix} v_1 \; v_2 \; ... \; v_k \end{bmatrix} \cdot \begin{bmatrix} Q_1 \\ Q_2 \\ ... \\ Q_k \end{bmatrix} \tag{5}$$

where w is a vector of models weights, S is a vector of single models scores, v is a vector of QMF weights and Q is a vector of QMF values.

Results

Model	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H		VoxSRC22 Dev	
	EER[%]	$DCF_{0.01}$	EER[%]	$DCF_{0.01}$	EER[%]	$DCF_{0.01}$	EER[%]	$DCF_{0.05}$
RC1	0.47	0.036	0.63	0.067	1.17	0.114	1.57	0.100
RC2	0.45	0.039	0.65	0.069	1.19	0.116	1.62	0.099
RC3	0.45	0.038	0.59	0.062	1.12	0.111	1.56	0.090
RC1-FT	0.44	0.030	0.56	0.063	1.07	0.105	1.45	0.089
RC3-FT	0.43	0.032	0.53	0.058	1.04	0.105	1.47	0.083
RC2-FT	0.36	0.037	0.55	0.060	1.05	0.104	1.42	0.088
SO1-FT	0.56	0.089	0.60	0.066	1.36	0.139	1.89	0.121
SO2-FT	0.49	0.071	0.59	0.071	1.30	0.135	1.68	0.108
RO1	0.34	0.020	0.48	0.047	0.85	0.076	1.25	0.068
RO2-FT2	0.20	0.012	0.42	0.041	0.80	0.076	1.16	0.065
RO2-FT1	0.20	0.014	0.45	0.043	0.89	0.080	1.29	0.072
RO2-FT3	0.20	0.017	0.42	0.040	0.80	0.076	1.15	0.066
RO1-FT	0.29	0.024	0.45	0.045	0.84	0.076	1.24	0.068
RO3-FT	0.14	0.019	**0.33**	**0.035**	0.68	0.063	0.96	0.059
RO4-FT	**0.13**	**0.011**	0.36	**0.035**	0.68	0.061	0.97	0.060
Fusion Closed	0.35	0.036	0.53	0.056	1.02	0.100	1.33	0.083
Fusion Open	0.14	0.012	0.36	**0.035**	**0.66**	**0.060**	**0.94**	**0.056**

Conclusions

- We have found out a **significant importance** of usage of QMF values in fusion.
- We also observed a **positive trend** in extending the amount of training speech data for open Track 2, as our **ResNet202** trained on a mixture of VoxCeleb2-dev and Self-VoxCeleb achieves state-of-the-art performance on the VoxCeleb1-test protocols.
- As a future work we would like to reach the supervised models quality with our **SSL** based models. We would also like to pre-train SSL models using a mixture of VoxCeleb2-dev and **Self-VoxCeleb** datasets.

ID R&D System Description to VoxCeleb Speaker Recognition Challenge 2022

Rostislav Makarov, Nikita Torgashov, Alexander Alenin, Ivan Yakovlev, Anton Okhotnikov

September 22, 2022

ID R&D Inc., USA, New York
{makarov,torgashov,alenin,yakovlev,ohotnikov}@idrnd.net