The ID R&D VoxCeleb Speaker Recognition Challenge 2023 System Description

Nikita Torgashov, Rostislav Makarov, Ivan Yakovlev,
Pavel Malov, Andrei Balykin, Anton Okhotnikov

August 20, 2023

ID R&D Inc., USA, New York
{torgashov,makarov,yakovlev,pavel.malov,andrew.balykin,ohotnikov}@idrnd.net

Overview

Datasets

- VoxTube-Large
- Domain Dataset Filtering

Architectures

Training

- Pre-training datasets
- Fine-tuning datasets

Scoring

- Scoring & Quality Measurement Functions
- Fusion scheme

Results

Conclusions

Datasets

- **Training data**
  - VoxTube-Large
  - VoxCeleb2
- **Validation data**
  - VoxCeleb1
  - VoxSRC-20, 21, 22, and 23 Dev
- **Augmentation**
  - MUSAN
  - Real RIRs

VoxTube-Large

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

The dataset size overcomes the VoxCeleb2 dataset size by a **multiple factor**, and contains about **100K** unique speakers.

A subset of VoxTube-Large is **open-sourced** and will be presented at the Speaker Recognition I section on Tuesday.

Domain Dataset Filtering

Not all speakers in the VoxTube-Large are equally important, due to the high domain gap.

Proposed algorithm:

- Extracted median embeddings for all speakers in VoxTube-Large and VoxCeleb1
- Identified top-50 most similar speakers from VoxTube-Large for each speaker in VoxCeleb1
- Removed speakers from VoxTube-Large with a cosine similarity greater than 0.8

This resulted in a refined domain subset - **VoxTube-30K**, which is 30% of the total dataset size.

Architectures

Model architectures:

- fwSE-ResNet100
- 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**, **XLSR**, and **UniSpeech**.

Training

Setup for the fwSE-ResNet100 model

- **Pre-training**
  - 4-second segments
  - 300 epochs
  - Data augmentation
- **Fine-tuning**
  - 6-second segments
  - 30 epochs
  - No data augmentation

Pre-training datasets

VoxSRC-23 validation

Pretrain dataset	Fine-tune dataset	EER,%	MinDCF
VoxTube-Large + VoxCeleb2	VoxTube-Large + VoxCeleb2	2.54	0.141
VoxTube-Large	VoxTube-Large + VoxCeleb2	**2.18**	**0.123**

Pretraining on **VoxTube-Large only** shows better performance, considering the same fine-tuning dataset.

Fine-tuning datasets

VoxSRC-23 validation

Pretrain dataset	**Fine-tune dataset**	**EER,%**	**MinDCF**
VoxTube-Large	VoxCeleb2 + VoxTube-Large	2.18	0.123
VoxTube-Large	VoxCeleb2 + VoxTube-30K	**1.94**	**0.105**

Fine-tuning on the **VoxTube-30K** works much better, compared to the fine-tuning on the full VoxTube-Large.

Scoring & Quality Measurement Functions

- **Scoring**
  - Cosine pairwise, $10 \times 4sec$ crops
  - AS-Norm, VoxCeleb2, top N=100
- **Audio content attributes**
  - Age
  - Gender
  - Speech length
  - Voice liveness score

- **Audio quality measurement**
  - NISQA model
  - Signal to Noise detector
  - Babble noise detector
- **Model embedding statistics**
  - L1 and L2 norm of embedding
  - Mean and STD of embedding

Fusion scheme

The output of our system is a linear fusion of normalized model scores and QMF values. To find the weights of each component in a **score-level** fusion we used a **Logistic Regression** model with a high L1 penalty on the **VoxSRC-23 dev** set.

Results

Model	Dataset	VoxSRC-23 Dev		VoxSRC-23 Eval	
		EER[%]	$DCF_{0.05}$	EER[%]	$DCF_{0.05}$
WavLM + ECAPA	VC2	3.64	0.195	-	-
WavLM + ECAPA	VC2 + VTL	**2.71**	**0.157**	-	-
fwSE-ResNet100	VC2	3.24	0.174	-	-
fwSE-ResNet100	VTL	2.73	0.156	-	-
fwSE-ResNet100	VC2 + VTL	**1.94**	**0.105**	2.14	0.110
Fusion	VC2 + VTL	1.45	0.086	1.88	0.096
Fusion + emb. QMF	VC2 + VTL	1.06	0.069	1.38	0.078
Fusion + all QMF	VC2 + VTL	**0.94**	**0.056**	**1.30**	**0.076**

Where *VC2* - VoxCeleb2, *VTL* - VoxTube-Large, *Fusion* - linear fusion of 10 models with AS-Norm

Conclusions

- The usage of **additional speech data** gives a significant performance boost
- The **domain dataset filtering** extracts the most useful part of the dataset
- The **embedding QMF** values play a crucial role in the fusion

The ID R&D VoxCeleb Speaker Recognition Challenge 2023 System Description

Nikita Torgashov, Rostislav Makarov, Ivan Yakovlev,
Pavel Malov, Andrei Balykin, Anton Okhotnikov

August 20, 2023

ID R&D Inc., USA, New York
{torgashov,makarov,yakovlev,pavel.malov,andrew.balykin,ohotnikov}@idrnd.net