VERIDAS SYSTEM DESCRIPTION FOR NIST SRE20 AND SRE21 CHALLENGES

Guillermo Barbadillo, Álvaro Martínez and Santiago Prieto

Veridas

ABSTRACT

In this report, we describe the submission of Veridas Digital Authentication Solutions S.L. team for the NIST 2020 CTS Speaker Recognition Challenge and for NIST 2021 Speaker Recognition Evaluation.

***Index Terms*—** speaker recognition, long duration, Convolutional Neural Network (CNN), telephone audio

1. INTRODUCTION

In this report, we describe the submission of Veridas Digital Authentication Solutions S.L. team for the NIST 2020 CTS Speaker Recognition Challenge and for NIST 2021 Speaker Recognition Evaluation.

Both submissions share many of the steps because they were developed at the same time. This document highlights the differences between both submissions that can be explained by the subtle differences in the evaluation datasets.

2. EXPERIMENTAL SETUP

2.1 Audio models pretrain on Voxceleb

The first step in the solution was to train the models on the Voxceleb dataset. The dataset was augmented using telephone codecs. This allowed to train the model both on an Afv (Audio from video) and a synthetic telephonic audio sets.

2.2 Audio models fine-tuning on datasets from the domain

The second step is to fine-tune the models using data that was closer to the domain of the challenge.

2.2.1 SRE20 fine-tuning

For SRE20 challenge the models were fine-tuned using SRE16 and SRE18 Evaluation and Development sets.

2.2.2 SRE21 fine-tuning

For SRE21 challenge the models were fine-tuned using NIST CTS superset and SRE16 Evaluation set.

2.3 Visual data

The visual model was trained with the VGG 2 [7] and a private dataset owned by Veridas.

2.4 Validation data

For SRE20 the public leaderboard was used as the validation set since it was possible to make up to 3 submissions a day.

On the other hand for SRE21 there was no public leaderboard, so the 2021 NIST SRE Development Set was used for validation.

2.5 Data augmentation

Standard data augmentation techniques were applied using audios from Musan [1] dataset and reverberations from RIRs [2] dataset.

2.6 Input features

All the models in the final system used Log Mel features as input. The number of filters was modified between 30 and 50 to induce variability between models.
Audio duration used as input for training was also different for the models to induce variability and values between 2 and 12 seconds were used.
Silence was removed from the audios using an energy-based Voice Activity Detector.

3. MODELS

3.1 ResNet

When it came to audio processing models, all the models used on the final solution were based on ResNet architecture as described in the BUT solution to VoxCeleb challenge [3], however instead of using ResNet34, the bigger ResNet152 was used.

In the case of the visual processing models, a single ResNet101 architecture was used as a backbone.

3.2 Additive angular margin loss

As proposed in the paper ArcFace: Additive Angular Margin Loss for Deep Face Recognition [4], m2 margin was used in all the models.

Different values of m2 between 0.2 and 0.5 were used to induce variability on the models of the ensemble.

3.3 Cyclic learning rates

Both on pretraining and on fine-tuning cyclic learning rates [5] were used for training the models.

3.4 Large Margin Cosine Loss

The visual model holds the particularity of being trained with the loss function proposed in CosFace: Large Margin Cosine Loss for Deep Face Recognition [8].

4. BACKEND

4.1 Euclidean distance

The embeddings of the speakers are restricted to lie on a hypersphere of radius 1. Following that restriction, using euclidean distance is equivalent to cosine distance but euclidean distance is faster to compute. Thus, we used euclidean distance for measuring similarity between embeddings.

There was no preprocessing nor centering of the embeddings.

4.2 Score normalization

We used adaptative symmetric score normalization (adapt S-norm) [6] with 250 top scoring speakers. The cohort was created using all training speakers.

4.3 Source type, language and audio duration calibration

The SRE21 challenge included comparisons between speakers of different source types (telephone and video) and also comparisons between different languages. On the development set it was observed that the distribution of the scores was slightly different based on those features.

Thus it was possible to improve the cost function by doing a simple linear calibration based on source type match and language match features.

However those features were not provided for the test set. By using audio extension (.sph or .flac) was possible to generate source type match features. For estimating the language match features a logistic regression model was built on top of biometric embeddings by training on the development set.

Finally it was also observed a drift on the scores due to the duration of the audios used on the comparison and this was also addressed with a linear calibration.

For SRE20 challenge only audio duration calibration was applied since it did not have different source or different languages comparisons.

4.4 Audio and Visual fusion

The fusion of voice and face biometric scores was done using a weighted average.

5. RESULTS

In the case of the results of NIST 2020 CTS Speaker Recognition Challenge, since the challenge is still ongoing, only our current results at the time of writing can be registered.

In the case of the NIST 2021 Speaker Recognition Evaluation, our results over the development set are provided because results on the evaluation set were not public at the time of writing. We don't differentiate between actual and min cost because we used the development for calibration and thus they are equal on that dataset.

5.1 NIST SRE20

Table 1: *Results on the leaderboard at the time of writing*

EER	Cost	model
2.6	0.110	Ensemble of 6 models

5.2 NIST SRE21 - Audio only

Table 2: *Results of the different models on dev audio only*

EER	Cost	model
8.5	0.435	Best single model without calibration
7.4	0.409	+ Source type match calibration
6.9	0.383	+ Language match calibration
6.9	0.381	+ Audio duration calibration
6.5	0.364	Ensemble of 6 models

5.3 NIST SRE21 - Visual only

Table 3: *Results of the different models on dev visual only*

EER	Cost	model
3.1	0.051	Best single model

5.4 NIST SRE21 - Audio-visual

Table 4: *Results of the different models on dev audio-visual*

EER	Cost	model
9.9	0.507	Best single model without calibration
8.8	0.416	+ Source type match calibration
8.3	0.388	+ Language match calibration
8.4	0.386	+ Audio duration calibration
8.13	0.381	Ensemble of 6 models
3.1	0.051	Face biometrics model
1.45	0.035	Fusion with face biometrics

6. PROCESSING TIME

It has been estimated that processing a single trial would cost less than a second using the Veridas cloud API.

7. REFERENCES

[1] MUSAN, http://www.openslr.org/17

[2] RIR_NOISES, http://www.openslr.org/28

[3] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, Oldřich Plchot "BUT System Description to VoxCeleb Speaker Recognition Challenge 2019"

[4] Jiankang Deng, Jia Guo, Niannan Xue, Stefanos Zafeiriou "ArcFace: Additive Angular Margin Loss for Deep Face Recognition"

[5] Leslie N. Smit, "Cyclical Learning Rates for Training Neural Networks" 2017

[6] Pavel Matejka, Ond ˇ rej Novotn ˇ y, Old ´ rich Plchot, Luk ˇ a´s Burget, ˇMireia Diez Sanchez, Jan "Honza" ´ Cernock ˇ y´ "Analysis of Score Normalization in Multilingual Speaker Recognition"

[7] Qiong Cao, et al. "VGGFace2: A dataset for recognising faces across pose and age." *13th IEEE International Conference on Automatic Face & Gesture Recognition*. 2018.

[8] Wang, Hao, et al. "Cosface: Large margin cosine loss for deep face recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.