Team 24 System Description for the SdSV Challenge

Jaun M. Coria, Hervé Bredin, Sahar Ghannay, Sophie Rosset

Université Paris-Saclay, CNRS, LIMSI

{coria, bredin, ghannay, rosset}@limsi.fr

1. Introduction

During the Short-duration Speaker Verification challenge [1], our team (Team 24) participated to Task 2: Text-Independent Speaker Verification. This paper describes the system of our official submission.

2. Architecture

The network architecture used combines SincNet trainable feature extraction [2] with the standard x-vector architecture [3] to build a fully end-to-end speaker verification system. Both SincNet and x-vector use the configuration proposed in the original papers (except for the SincConv layer of SincNet that uses a stride of 5 for efficiency).

As depicted in Figure 1, the network takes the waveform as input and returns 512-dimensional speaker embedding. In practice, we use a 3s-long sliding window with a 100ms step to extract a sequence of speaker embeddings that are then averaged to obtain just one speaker embedding per file. These average speaker embeddings are then simply compared with the cosine distance.

3. Additive angular margin loss

The cross entropy loss $\mathcal{L}_{\text{CE}}$, initially introduced for multi-class classification, is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{\exp(\sigma_{iy_i})}{\sum_{k=1}^{K}\exp(\sigma_{ik})}\right] \tag{1}$$

where N is the number of training examples (here, audio segments x_i), K the number of classes (here, speakers) in the training set, y_i is the class of training sample x_i, and σ_i is the output of a linear classification layer with weights $C \in \mathbb{R}^{m\times K}$ and bias $b \in \mathbb{R}^K$:

$$\sigma_i = f(x_i)\cdot C^T + b \tag{2}$$

Equation 2 can be rewritten as follows:

$$\forall k \;\; \sigma_{ik} = \|f(x_i)\|\cdot\|c_k\|\cdot\cos\theta_{ic_k} + b_k \tag{3}$$

where θ_{ic_k} is the angular distance between the representation $f(x_i)$ of training sample x_i, and c_k the k$^{\text{th}}$ row of matrix C.

The additive angular margin loss [4] normalizes row vectors c_k and representation $f(x_i)$, and introduces a margin to penalize the angular distance between a representation $f(x_i)$ and its center c_{y_i}:

$$\forall k \;\; \sigma_{ik} = \begin{cases}\alpha\cdot\cos(\theta_{ic_k}+m) & \text{if } y_i = k\\ \alpha\cdot\cos\theta_{ic_k} & \text{otherwise}\end{cases} \tag{4}$$

where the kth row of matrix C can be seen as a canonical representation of the kth speaker, m is the margin and α scales the cosine. This loss explicitly forces embeddings to be closer to their centers by artificially augmenting their distance by the margin.

4. Training

The official training set [5] was split into Train, consisting of 488 random speakers, and Dev with the remaining 100 speakers.

The model was pretrained for 560 epochs on VoxCeleb 2, and then fine-tuned on the Train split until convergence (validated on the Dev split), which happened after 5 epochs. Both these training runs benefited from on-the-fly background noise augmentation from the MUSAN database [6] and were optimized using the additive angular margin loss.

5. Results

Official evaluation consisted of the minimum detection cost function (minDCF) as stated in the evaluation plan [1]. A *progress* set for evaluation was available during the model development period, while a final *evaluation* set was released afterwards. Detailed results on both sets are summarized in Table 1, while the DET curves of our model and the baseline are shown in Figure 2.

	progress EER	progress minDCF	evaluation EER	evaluation minDCF
overall	5.98	0.264	5.96	0.265
male	4.89	0.222	4.92	0.225
female	6.31	0.277	6.26	0.277
EN	6.72	0.299	6.68	0.300
FA	5.34	0.237	5.33	0.237
EN male	5.22	0.247	5.26	0.253
EN female	7.02	0.313	7.02	0.313
FA male	4.60	0.204	4.63	0.205
FA female	5.57	0.247	5.53	0.247
TC vs IC	7.58	0.271	7.50	0.272

Table 1: *Our team's results on the progress and evaluation sets under different constraints*

6. Conclusion

The system described here was part of a greater work on comparing loss functions for end-to-end speaker verification [7]. The code needed to run our experiments, as well as the pretrained model on VoxCeleb 2 are available as open source[1].

7. References

[1] K. A. Zeinali, Hossein nad Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SdSV) challenge 2020: the challenge evaluation plan." arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[1] github.com/juanmc2005/SpeakerEmbeddingLossComparison

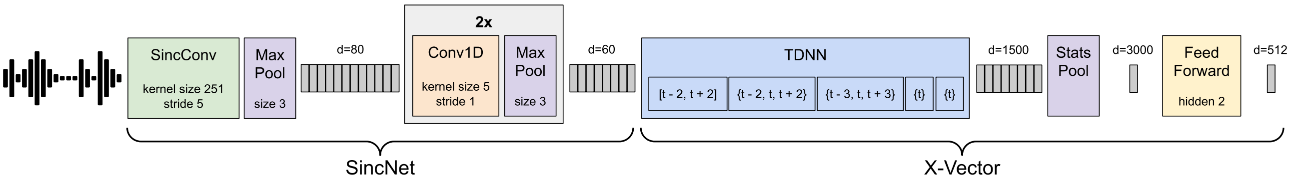

Figure 1: *The end-to-end architecture combines SincNet trainable features with the standard TDNN x-vector architecture.*

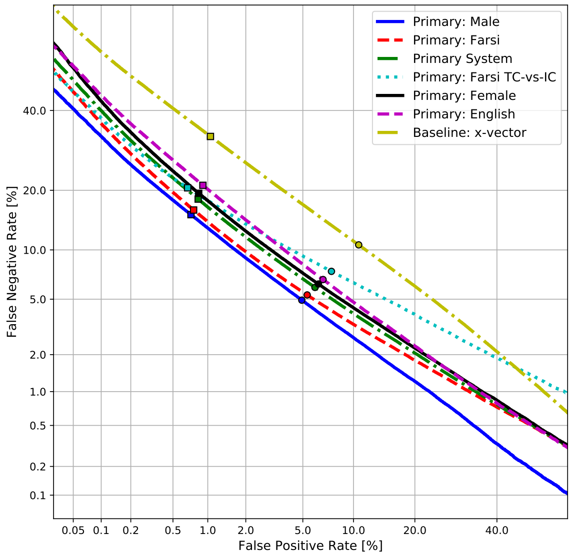

Figure 2: *DET curve of our system (Primary) on different subconditions compared to the x-vector baseline.* ○ *corresponds to the EER, while* ◇ *corresponds to the minDCF*

[2] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[4] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019.

[5] H. Zeinali, L. Burget, and J. Cernocky, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database," in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[6] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[7] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification," 2020.