Fine-Tuning SimAM-ResNet34 and WavLM-Base for Cross-Lingual Speaker Verification

Araj Shah
Howard University, araj.shah@bison.howard.edu
Howard Prioleau
Howard University, utsav.shah@bison.howard.edu
Dr. Saurav Aryal
Howard University, saurav.aryal@howard.edu
Dr. Gloria Washington
Howard University, gloria.washington@howard.edu

We present a lightweight, reproducible submission for the TidyVoice 2026 cross-lingual speaker verification challenge implemented in the WeSpeaker toolkit under single-GPU Google Colab constraints. Our primary system, S1, uses the official SimAM-ResNet34 checkpoint pretrained on VoxBlink2 and VoxCeleb2 and fine-tuned on TidyVoiceX, which we further fine-tune for five epochs with large-margin classification. In parallel, we implement a secondary self-supervised system, S2, using a frozen WavLM-Base frontend with a compact statistics pooling speaker head, trained for four epochs. Both systems use standard speech augmentation during training with MUSAN noise and RIRS reverberation, while inference uses clean embeddings and cosine scoring. To combine systems, we perform score-level fusion calibrated on a labeled Tune-S development split. We z-normalize each system's Tune-S scores using their mean and standard deviation, grid-search a convex fusion weight alpha in the range 0 to 1 with step 0.01 to minimize EER, and apply the frozen normalization and alpha to fuse Eval-A (Task 1) and Eval-U (Task 2) score files for submission. On Tune-S, S1 substantially outperforms S2, so the selected fusion weight is alpha equals 1.0.

CCS CONCEPTS • **Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI•Computing methodologies~Machine learning~Machine learning approaches** • **Computing methodologies~Machine learning~Machine learning approaches~Neural networks**

Additional Keywords and Phrases: cross-lingual speaker verification, SimAM-ResNet34, WavLM-Base self-supervised features, cosine similarity scoring, score-level fusion, z-score normalization (z-norm), MUSAN and RIRS augmentation, reproducible WeSpeaker pipeline

REFERENCES

[1] Yang, L., Zhang, R.-Y., Li, L., & Xie, X. (2021). SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. ICML.

[2] Miao, Z., Li, M., Wang, W., Zhang, C., Li, C., Li, J., Zhang, M., & Xiao, X. (2022). WeSpeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit. (arXiv:2210.17016)

[3] Chen, S., Wu, C., Wang, Z., Chen, Z., Xu, J., Yao, Z., Liu, S., et al. (2021). WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. (arXiv:2110.13900)

[4] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep Speaker Recognition. Interspeech. (arXiv:1806.05622)

[5] Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A Music, Speech, and Noise Corpus. (arXiv:1510.08484)

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. CVPR. (arXiv:1512.03385)

[7] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. CVPR. (arXiv:1801.07698)

[8] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). The Kaldi Speech Recognition Toolkit. IEEE ASRU.