TECHNICAL REPORT SP CUP 2024 - TEAM IITH

Atharv Ramesh Nair [1] *Anirudh Srinivasan* [2] *Tejadhith S* [1] *Vaideeswaran A.P* [1]
Himanshu Kumar Gupta [3] *Sreekanth Sankala* [1] *K. Sri Rama Murthy* [1]

[1] Department of Electrical Engineering, Indian Institute of Technology Hyderabad
[2] Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad
[3] Department of Artificial Intelligence, Indian Institute of Technology Hyderabad

ABSTRACT

This study explores text-independent far-field speaker recognition, emphasizing challenges highlighted in the Robovox challenge, particularly in noisy and echoic environments. It evaluates state-of-the-art methods and results show the superiority of the ERes2Net model trained on the 3D speaker dataset. Detailed insights into datasets, architectures, and training protocols are provided, showcasing ERes2Net's ability to handle local and global functions. Our most successful model achieves a minimum Detection Cost Function (min DCF) of 0.7096 on the Robovox Dataset. Furthermore, through model ensembling, we manage to decrease the Equal Error Rate (EER) to 9.59%.

1. INTRODUCTION

In today's technological landscape, there is a growing demand for reliable speaker recognition systems, particularly in fields like security, human-robot interaction, and personalization services. The Robovox challenge focuses on text-independent far-field speaker verification by a mobile robot a noisy and reverberating environment. The competition poses a lot of challenges including signal attenuation due to ambient noise, additive noise, reverberation, babble noise, and non-stationary channel characteristics resulting from varying recording distances. Traditional speaker models trained on clean data struggle in such conditions, necessitating novel techniques.

Various methods have been proposed to address reverberation and noise challenges in far-field scenarios for Automatic Speaker Verification (ASV) systems. Signal-level techniques like weighted prediction error [1, 2] aid in dereverberation, while DNN-based denoising [3, 4, 5] and beamforming [6, 7] enhance speech quality in single-channel and multi-channel setups, respectively. At the modeling level, strategies such as data augmentation [8, 9] and transfer learning [10] are effective with limited data. Adversarial training [11, 12] and variability-invariant loss [13] help learn noise-invariant speaker embeddings. Joint training of speech enhancement and speaker embedding networks boosts ASV performance in noise [14, 15, 16]. A multichannel training framework improves deep speaker modeling with microphone arrays [17], and enrollment data augmentation minimizes mismatch between enrollment and testing utterances [10].

We conducted experiments using state of the art Deep Neural Network architectures including ECAPA-TDNN [18], ResNetSE34v2 [18], wavLM (self-supervised)[19], and ERes2Net [20]. These models were evaluated using the VoxCeleb [21], Common Voices [22] and 3D-Speaker [23] Datasets. After thorough experimentation, we determined that the ERes2Net (Large) model, trained on the 3D-Speaker Dataset, outperformed the others. The rest of this report is structured as follows: Section 2 provides detailed descriptions of the RoboVox and 3D-Speaker Datasets. Section 3 delves into the intricacies of the ERes2Net Architecture. In Section 4, we present our system specifications and training procedures. Section 5 outlines the results obtained, and we conclude the report by discussing future plans in Section 6.

2. DATASETS

2.1. Challenge Dataset

In this challenge, a novel benchmark is introduced to advance research in far-field single-channel and multi-channel speaker verification. The evaluation benchmark utilizes the Robovox French corpus, recorded by a mobile robot equipped with a speaker recognition system in diverse acoustic conditions. The robot has three external microphones and one embedded microphone (Channel 4). A ground truth microphone (Channel 5) is placed near the speaker's mouth. The dataset comprises 78 speakers engaged in 2219 conversations, with an average of 5 dialogues per conversation. Each dialogue is approximately 3.6 seconds long.

Notably, Channel 5 provides a clean signal for establishing a baseline system. As part of the single channel track, we are supposed to use Channel 5 for Enrollment and Channel 4 for test data. The dataset also incorporates various distances (1m, 2m, 3m) and acoustical environments (hall, open space,

Thanks to the Speech Information Processing Lab, IITH for computational resources.

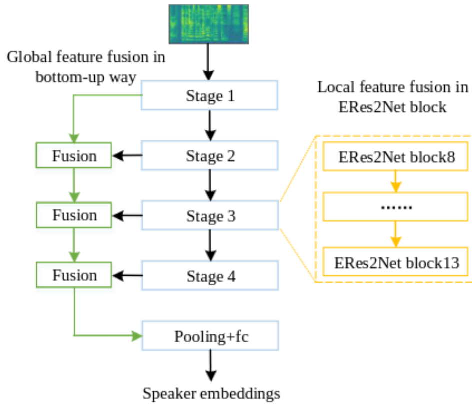

Fig. 1. Overview of ERes2Net Architecture

small room, medium room) with open or closed doors. Different robot placements (wall, center, corner) introduce challenges like severe reverberation. A file containing enrollment utterance and speaker utterances have been provided and we need to generate similarity scores. Minimum Detection Cost Function (Min DCF) and Equal Error Rates (EER) have been used as primary and secondary metrics respectively.

2.2. Training Dataset

We opted to utilize the 3D-Speaker Data [23] for model training. The primary rationale for this dataset selection stems from its comprehensive composition, encompassing a training dataset featuring 10,000 speakers and 579,013 utterances, with a cumulative valid speech duration of 1124 hours. Significantly, the dataset includes speech recordings obtained at varying distances, ranging from 0.1m to 4m, thus providing a diverse set of far-field speech data for robust model training.

Noteworthy attributes of the dataset include its incorporation of recordings in 14 distinct Chinese dialects, captured using different recording devices. This linguistic and acoustic diversity is particularly advantageous for addressing challenges associated with out-of-domain data. Specifically, the inclusion of 14 Mandarin dialects enriches the dataset, contributing to the model's ability to handle linguistic variations inherent in diverse speech datasets.

3. MODEL ARCHITECTURE

The ERes2Net architecture is an extension of the Res2Net model, designed to overcome limitations in local information interaction and global perspective. It includes two Local Feature Fusion (LFF) branch and a Global Feature Fusion (GFF) branch.

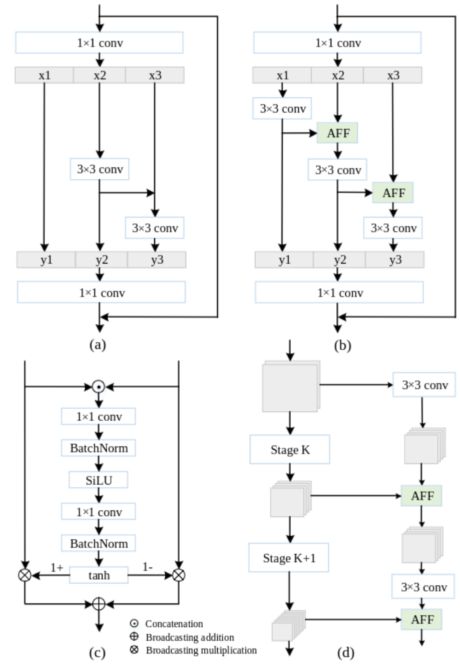

Fig. 2. (a) Res2Net block (b) ERes2Net block (c)Attentional feature fu- sion (AFF) module; (d) Global feature fusion (GFF) module.

3.1. *Local Feature Fusion (LFF)*

The LFF block incorporates an attentional feature fusion (AFF) mechanism to enhance fine-grained features and promote local information interaction. Feature maps are organized into groups, and the AFF module calculates local attention weights for adjacent feature maps. The hierarchical fusion structure within the LFF block expands the model's receptive fields, enabling improved integration of local information across different channels.

3.2. *Global Feature Fusion (GFF)*

The GFF component focuses on augmenting global feature interaction, particularly in the bottom-up pathway. Multi-scale features from each ERes2Net stage undergo down-sampling, and attention weights are computed using the AFF module. Down-sampled feature maps are then modulated through bottom-up attention, enhancing the model's ability to capture features at various temporal scales.

The ERes2Net architecture facilitates the extraction of both local and global patterns in input signals, contributing to heightened accuracy and robustness in speaker verification systems

Model	5-5	5-4
wavLM	6.16	14.6
ECAPA	7.14	14.2
SEResNet34V2	7.2	14.1
ERes2Net	**5.9**	**11.2**

Table 1. *EER Performance of different models on Robovox -Multi Channel Data*

Enrollment-Test	Common	VoxCeleb	3D base	3D large
5-5	6.3	5.9	7.5	**8.08**
5-4	12.3	12.8	11.6	**11.3**

Table 2. *EER Performance of ERes2Net trained on different Datasets*

4. SYSTEM SPECIFICATIONS

Our system closely follows the original ERes2Net framework [20]. The acoustic features utilized in the study consist of 80-dimensional Filter Bank (FBank) representations, computed with 25ms windows and a 10ms shift. During the training phase, 3-second segments are randomly cropped from each utterance. Data Augmentation techniques include the incorporation of RIR (Room Impulse Response) and Musan (additive noise), obtained from the 3D-Speaker Dataset [23].

Stochastic gradient descent (SGD) optimizer is employed, accompanied by a cosine annealing scheduler and a linear warm-up scheduler. Initially, over the first 5 epochs, the learning rate gradually increases to 0.2. Momentum value was set to 0.9, weight decay to 10^{-4}, Angular Additive Margin Softmax (AAM-Softmax) [22] was used as the loss function for training . Speaker embeddings, of dimensionality 512, are extracted from the first fully-connected layers of the model. Speed perturbation is incorporated during training by introducing factors of 0.9, 1.0, and 1.1 with equal probabilities. Cosine Similarity is applied to mean-subtracted and unit length-normalized embeddings to obtain similarity scores for the backend. During evaluation on the Challenge Dataset, embeddings are extracted by randomly cropping fixed length segments. All training was done using the PyTorch framework .

5. EXPERIMENTS AND RESULTS

We utilized the Multi-Channel Data as our validation set. Initially, we employed the Equal Error Rate (EER) as a validation metric to identify the best models and datasets. Table 1 presents the performance of different pretrained models on Robovox multichannel data. We computed EERs for 5-5 and 5-4 Enrollment and Test Channel Pairs. Notably, the results demonstrate that ERes2Net outperforms other models in both tasks. Further details are provided in Table 2, illustrating the performance of ERes2Net pretrained on various datasets. It is evident that the model trained on the 3D-Speaker Dataset exhibits superior performance.

	1s	2s	3s	4s	5s	Original Length
EER (%)	12.88	12.06	11.67	11.23	**10.97**	11.09
MinDCF	0.64	0.56	0.52	0.51	**0.505**	0.506

Table 3. *Multi-Channel (5v4) Performance on varying length test utterances*

Model	EER	Min DCF
ECAPA-TDNN + ERes2Net (Ensemble)	**9.59**	0.7203
ERes2Net	10.59	**0.7059**

Table 4. *Models with best Performance*

Moreover, Table 3 provides both the EER and Min-DCF metrics while varying the length of the randomly cropped test utterance. These metrics were obtained without centering and normalizations initially, with a significant drop observed in DCF after mean subtraction and unit length normalization. Based on these results, we concluded to focus on the ERes2Net model and utilize the 3D-Speaker (Large) dataset for training.

We also explored extracting the Room Impulse Response (weiner filter based) and additive noise components using sample data (30 min), which were subsequently used for data augmentation during the training procedure. This didn't show much improvement. This could be because weiner filters may not be suitable for non-stationary conditions

Table 4 shows our best performing models. The ensemble model (ECAPA-TDNN+ERes2Net) gives the best EER wheras ERes2Net alone gives the best Min DCF. Ensembling of models led to increase in min DCF

6. CONCLUSION

Our study leveraged the ERes2Net Architecture, coupled with the 3D-Speaker Dataset, yielding competitive results. Initially, we achieved a baseline Equal Error Rate (EER) of 5.9% when evaluating Multichannel data with Enrollment and Test Channels - 5, aiming for similar performance with Channel 4 for testing. Implementing more robust data augmentation during training could enhance model robustness against various types of noise, offering a potential direction for future research. Additionally, exploring meta-learning techniques for short utterance speaker recognition with imbalance length pairs, inspired by Kye et al. [24], may prove useful. Utilizing Prototypical Networks trained with support sets of long utterances and query sets of short utterances could significantly improve speaker recognition performance.

7. REFERENCES

[1] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, pp. 2707–2720, 12 2012.

[2] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1717–1731, 2010.

[3] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[4] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," 2017.

[5] Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," 01 2017, pp. 136–140.

[6] Xiaoyi Qin, Hui Bu, and Ming Li, "Hi-mia : A far-field text-dependent speaker verification database and the baselines," 12 2019.

[7] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, p. 5554–5558, IEEE Press.

[8] Danwei Cai, Xiaoyi Qin, Weicheng Cai, and Ming Li, "The DKU System for the Speaker Recognition Task of the 2019 VOiCES from a Distance Challenge," in *Proc. Interspeech 2019*, 2019, pp. 2493–2497.

[9] Pavel Matejka, Oldrich Plchot, Hossein Zeinali, Ladislav Mosner, Anna Silnova, Lukás Burget, Ondrej Novotný, and Ondrej Glembek, "Analysis of but submission in far-field scenarios of voices 2019 challenge," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, Gernot Kubin and Zdravko Kacic, Eds. 2019, ISCA.

[10] Xiaoyi Qin, Danwei Cai, and Ming Li, "Far-Field End-to-End Text-Dependent Speaker Verification Based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation," in *Proc. Interspeech 2019*, 2019, pp. 4045–4049.

[11] Ruiteng Zhang, Jianguo Wei, Xugang Lu, Wenhuan Lu, Di Jin, Lin Zhang, Yantao Ji, and Junhai Xu, "Self-supervised learning based domain regularization for mask-wearing speaker verification," *Speech Commun.*, vol. 152, no. C, jul 2023.

[12] Kihyun Nam, Youkyum Kim, Jaesung Huh, Hee Soo Heo, Jee weon Jung, and Joon Son Chung, "Disentangled representation learning for multilingual speaker recognition," 2023.

[13] Danwei Cai, Weicheng Cai, and Ming Li, "Within-sample variability-invariant loss for robust speaker recognition under noisy environments," 2020.

[14] Jianchen Li, Jiqing Han, and Hongwei Song, "Gradient regularization for noise-robust speaker verification," 08 2021, pp. 1074–1078.

[15] Suwon Shon, Hao Tang, and James Glass, "VoiceID Loss: Speech Enhancement for Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 2888–2892.

[16] Shuo Liu, Andreas Triantafyllopoulos, Zhao Ren, and Björn Schuller, "Towards speech robustness for acoustic scene classification," 10 2020.

[17] Danwei Cai, Xiaoyi Qin, and Ming Li, "Multi-Channel Training for End-to-End Speaker Recognition Under Reverberant and Noisy Environment," in *Proc. Interspeech 2019*, 2019, pp. 4365–4369.

[18] Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," .

[19] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[20] Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi, "An enhanced res2net with local and global feature fusion for speaker verification," 2023.

[21] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[22] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, July 2018.

[23] Yafeng Chen Hui Wang Siqi Zheng, Luyao Cheng and Qian Chen, "3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement," 2023.

[24] Seong Min Kye, Youngmoon Jung, Hae Beom Lee, Sung Ju Hwang, and Hoirin Kim, "Meta-learning for short utterance speaker recognition with imbalance length pairs," in *Interspeech*, 2020.