Enhancing Arabic Speaker Recognition with ECAPA-TDNN

Mahmoud Ayman
Research and Innovation Department, T2 Company, Riyadh, Saudi Arabia
m.sayed@t2.sa (corresponding author)

Fahad A. Aloufi
Department of Cybersecurity, College of Computer, Qassim University, Qassim, Saudi Arabia | Research and Innovation Department, T2 Company, Riyadh, Saudi Arabia
faa.alharbi@qu.edu.sa

Received: 6 August 2025 | Revised: 15 September 2025, 27 October 2025, 8 March 2026, 29 March 2026, 30 March 2026, and 2 April 2026 | Accepted: 3 April 2026

 DOI: https://doi.org/10.48084/etasr.13902

ABSTRACT

This paper presents a fine-tuned Emphasized Channel Attention, Propagation and Aggregation - Time Delay Neural Network (ECAPA-TDNN) model for Arabic speaker recognition, with a focus on enhancing performance in noisy environments. The model was trained on the Voice of Celebrities 1 (VoxCeleb1) and VoxCeleb2 corpora combined with Arabic data from the Qatar Computing Research Institute (QCRI) Aljazeera Speech Resource (QASR), and was evaluated on the VoxCeleb1 test protocol (Vox1-O), the Arab Celebrity (ArabCeleb) dataset, a held-out QASR test split, and an in-house Arabic dataset of authentic recordings. Through targeted fine-tuning and data augmentation techniques, the proposed approach reduces the Equal Error Rate (EER) on Arabic datasets and improves robustness to noise, while maintaining satisfactory performance on English datasets. These findings indicate that careful adaptation can support the development of more balanced multilingual speaker verification systems, particularly for underrepresented languages such as Arabic.

Keywords-ECAPA-TDNN; speaker verification; speaker embeddings; noise

I. INTRODUCTION & BACKGROUND

Speaker verification plays a crucial role in biometric authentication, where the objective is to verify an individual's identity based on voice characteristics. With the increasing adoption of voice-driven technologies, there is a growing demand for speaker verification systems that are both accurate and robust across different languages and acoustic environments. However, achieving reliable performance under diverse real-world conditions, including varying noise levels and language-specific phonetic characteristics, remains a significant challenge in this field.

The field of speaker verification has advanced considerably so far through methods such as x-vectors [1] and Emphasized Channel Attention, Propagation and Aggregation - Time Delay Neural Network (ECAPA-TDNN), which have improved accuracy across various acoustic conditions. Building on the success of the x-vector framework, the ECAPA-TDNN model [2] introduced enhancements including channel attention, Res2Net modules, and Squeeze-and-Excitation (SE) blocks, further improving verification accuracy, especially in noisy environments. These methods are commonly evaluated on Voice of Celebrities (VoxCeleb)-style protocols using the Equal Error Rate (EER) as the standard speaker verification metric [1-5], a convention also followed in this work.

In addition to these prominent architectures, other approaches have explored alternative strategies for improving speaker verification. For instance, authors in [6] employed TitaNet, which uses one-dimensional depth-wise separable convolutions and channel-attention-based pooling for speaker verification and diarization. In addition, authors in [7] utilized Context-Aware Masking++ (CAM++), focusing on achieving fast and efficient speaker verification, while authors in [8] employed Speech Neural Architecture Search (SpeechNAS), which optimizes TDNN frameworks through neural architecture search with emphasis on both efficiency and performance. Additionally, the extended U-Net model, proposed in [9], demonstrated improved speaker recognition in noisy environments by effectively handling diverse noise conditions. Another proposed model is SincNet, which processes raw waveforms using parametrized sinc functions and also improves speaker recognition accuracy by learning more meaningful filters [10]. Moreover, the complementary work/authors in [11] tackled noise robustness at the front-end by coupling Perceptual Wavelet Packet (PWP) thresholding with Mel-Frequency Cepstral Coefficients (MFCCs) and a Support Vector Machine (SVM) backend, achieving 99% real-

time Automatic Speech Recognition (ASR) accuracy on a low-power Raspberry Pi, and underscoring the effectiveness of wavelet-based denoising for edge deployments in noisy conditions.

The development of multilingual speaker verification systems has also received considerable scientific attention. While authors in [12] highlighted the difficulties associated with cross-modal verification across languages, authors in [13] proposed disentangled representation learning to separate language-specific and speaker-specific features, thereby improving multilingual performance. In addition, authors in [14] introduced a meta-learning approach for cross-channel verification that significantly improved accuracy under varying channel conditions [14].

Despite progress in multilingual speaker verification, Arabic-specific datasets are scarce, with limited benchmarks for evaluating Arabic speaker verification systems being available. This limitation, along with the challenging phonetic diversity of the Arabic language, has hindered the development of accurate Arabic speaker recognition systems. On top of that, models pre-trained on English corpora perform poorly when applied to linguistically different languages, highlighting the language dependency of speaker verification models [15]. Furthermore, even models trained on multilingual audio data without explicit language constraints, including ECAPA-TDNN, perform variably based on the language investigated. For instance, authors in [16] evaluated pre-trained models, including ECAPA-TDNN, on data from English, German, Danish, Spanish, and Arabic speakers, showing that the models performed well on the European languages, while for Arabic, the EER was 8.26% higher, further demonstrating that even multilingual models may struggle with languages that exhibit unique phonetic structures.

In an attempt to produce a speaker verification system for Arabic, this paper presents a fine-tuned ECAPA-TDNN model, leveraging the SpeechBrain framework to develop custom multilingual training recipes for Arabic speaker verification. ECAPA-TDNN was selected in this work because of several important advantages. Its SE blocks dynamically recalibrate channel-wise features to emphasize the most informative speaker characteristics, while its multi-scale feature extraction captures both local and global acoustic patterns, which is particularly important for handling the phonetic diversity of Arabic. In addition, its demonstrated robustness under noisy conditions and efficient computational performance make it well-suited for real-world applications. Meanwhile, the fine-tuning strategy enabled the model to learn Arabic-specific features while preserving the previously learned English representations, thus improving Arabic speaker verification performance without compromising English-language performance. Training data for the proposed model are drawn from the VoxCeleb1 and VoxCeleb2 corpora together with the Qatar Computing Research Institute (QCRI) Aljazeera Speech Resource (QASR) Arabic corpus, while evaluation is performed on four held-out test sets: the VoxCeleb1 (Vox1-O) protocol, the Arab Celebrity (ArabCeleb) evaluation files, a held-out QASR test split with unseen speakers, and an in-house Arabic dataset of authentic recordings collected by the T2 Company.

II. METHODOLOGY

A. Data

This study employed several datasets, both for training and testing the proposed fine-tuned ECAPA-TDNN model. For training, the VoxCeleb1 and VoxCeleb2 datasets [3-5] were employed, sourced from YouTube, which include over 1.24 million utterances from 7,205 speakers, totaling about 2,690 h of speech in English. In addition, for training, this study also employed the QASR dataset, which is a comprehensive Arabic speech corpus derived from Al Jazeera broadcasts, encompassing approximately 2,041 h and 1.6 million segments from 3,545 episodes. It features multi-layer annotations suitable for a variety of speech and language processing tasks, including speech recognition, dialect identification, and speaker identification. With 27,977 speakers and segments averaging 4 s, QASR is crucial for advancing Natural Language Processing (NLP) applications in Arabic [18]. From the publicly released QASR archive, 3,927 unique speakers were extracted; speakers with fewer than eight usable utterances after preprocessing or whose audio was unusable were excluded from this study's pool. The remaining audio was then cleaned and filtered with a Voice Activity Detection (VAD) filter from SpeechBrain's VAD-Convolutional Recurrent Deep Neural Network (CRDNN)-LibriParty 1 model, discarding non-speech regions and segments shorter than approximately 1 s so that only high-quality speech segments were retained. From this cleaned pool, this work then randomly selected approximately 60% of the speakers (2,502 speakers) for the training subset, while the remaining ~40% (1,425 speakers) were held out as unseen speakers for testing. Within each selected training speaker, the number of utterances was capped to balance the Arabic data and avoid over-representing speakers with very long broadcast appearances. To address the issue of catastrophic forgetting, the present study retained part of the original English data during fine-tuning. This strategy ensures that the model retains its ability to recognize English features while improving its performance on Arabic tasks. Table I summarizes the training data used in this study.

TABLE I. TRAINING DATA USED

Dataset	Total hours	Number of utterances	Number of speakers
VoxCeleb1 and 2	2,690	1.24M	7,205
QASR	~176	736,914	2,502

For the testing of the proposed model, the first benchmark employed was the Vox1-O evaluation from VoxCeleb1, involving 37,000 trials from 40 speakers [3-5]. The second evaluation dataset was the ArabCeleb [17], which features 1,930 utterances from 100 Arabic-speaking celebrities, extracted from YouTube videos, totaling 6 h of speech. Evaluation was performed using the official ArabCeleb speaker verification protocol distributed with the dataset, consisting of a fixed list of 8,096 verification trial pairs over 40 held-out speakers. The remaining speakers in the public dataset were not included in this protocol and were therefore excluded from

evaluation. This approach ensures direct comparability with previously reported ArabCeleb verification results.

Finally, an in-house dataset was employed that comprises 51 h of high-quality, authentic audio recorded from 706 users, totaling 86,285 files. This collection captures a diverse array of phrases in real-life scenarios with ambient noise, primarily from Arabic speakers with some English intermixed, making it ideal for robust speaker verification systems. The dataset was constructed in-house by the Research and Innovation Department at T2 Company specifically as an evaluation set for Arabic speaker verification. Audio was collected from native Arabic-speaking volunteers (employees and external contributors) recruited internally; participants were geographically distributed across Saudi Arabia and represented a mix of regional Arabic dialects. Each speaker was recorded across multiple sessions in everyday office and home environments using consumer-grade equipment (smartphone microphones and standard headsets) at a 16 kHz mono sampling rate. During each session, the speaker first read a set of predefined Arabic prompts (digits, short phrases, and longer sentences) and then engaged in spontaneous, free-form speech to capture realistic prosodic and acoustic variability, including ambient noise from fans, traffic, and background conversation. Recordings were segmented automatically and then manually screened to remove silent or corrupted files; only speakers with at least eight usable utterances were retained. No personal identifying information beyond an anonymous speaker ID was stored, and informed consent was obtained from every participant for research use.

For evaluation, the same quality-control criteria applied to QASR were also applied to the in-house dataset. Speakers whose recordings could not be reliably segmented, or who retained fewer than eight usable utterances after VAD-based cleaning, were excluded to ensure sufficient material for generating balanced positive (same-speaker) and negative (different-speaker) verification pairs. After filtering, 469 of the original 706 speakers were retained, generating the 1,288,110 evaluation pairs. For the QASR and in-house datasets, the custom evaluation files were generated with balanced numbers of positive (same-speaker) and negative (different-speaker) pairs. The evaluation protocol was text-independent, meaning that utterances within a comparison pair were not required to contain the same sentence or phrase. Speaker identity was determined exclusively from voice characteristics. Let x_i and x_j be utterances from the same speaker, and x_k be an utterance from a different speaker. Pairs were selected at random to ensure balanced coverage of speakers and conditions. Pair generation follows:

$$\text{Positive: } (x_i, x_j) \text{ with } s(x_i) = s(x_j) \quad (1)$$

$$\text{Negative: } (x_i, x_k) \text{ with } s(x_i) \neq s(x_k) \quad (2)$$

where $s(\cdot)$ returns the speaker identity. Table II summarizes the testing data used for evaluation.

1) Data Augmentation

To improve the model's robustness, data augmentation was applied during the training phase by introducing various types of noise, including music, speech babble, and environmental sounds, from the Music, Speech, and Noise (MUSAN) dataset [19]. The noise was added at random Signal-to-Noise Ratio (SNR) levels, simulating real-world noisy conditions. Additionally, speed perturbation was applied by randomly varying the speed of the speech signal, simulating different speaking rates. The augmented signal $y(t)$, where $x(t)$ represents the original audio signal and $n(t)$ the noise signal, is defined as:

$$y(t) = \frac{x(t) + \frac{n(t)}{\alpha}}{2} \quad (3)$$

where the scaling factor α is computed as:

$$\alpha = \frac{|x(t)|_2}{|n(t)|_2} \cdot 10^{\frac{\text{SNR}}{20}} \quad (4)$$

Different augmentation strategies were applied for the training and testing data. Specifically:

- For training: Noise from the MUSAN dataset (environmental sounds, speech babble, and music) was introduced at random SNR levels between 0 dB and 15 dB, while speed perturbation was also applied to simulate varying speaking rates.
- For testing: Pairs of original and augmented files were generated with noise at fixed SNR levels (0, 10, 20 dB), without speed perturbation.

TABLE II. TESTING DATA USED

Dataset	Number of pairs	Number of users
VoxCeleb1	37,000	40
ArabCeleb	8,096	40
QASR	28,262	1,425
In-house	1,288,110	469

B. Architecture

The ECAPA-TDNN architecture [2], implemented using the SpeechBrain framework [20], was adopted for the speaker verification process. The original model was trained using a softmax classifier with 7,205 output classes corresponding to the number of speakers in the VoxCeleb dataset. In this study, the softmax classifier was removed, and the 192-dimensional embedding layer was used to generate speaker embeddings. To enable multilingual speaker verification, the ECAPA-TDNN architecture was customized by expanding the classifier's output classes from 7,205 to 9,707, thus integrating Arabic speakers alongside the original English-based VoxCeleb dataset. In addition to the classifier modification, the model was fine-tuned using both English and Arabic data. The pretrained weights from the original model were retained during fine-tuning to leverage the previously learned features, while adjustments were made to the input layer to accommodate the phonetic differences between English and Arabic.

III. EXPERIMENTS

Experiments were conducted on a cost-efficient hardware setup, utilizing an AMD Ryzen 3950X Central Processing Unit (CPU) and an NVIDIA GeForce RTX 3080 Graphics Processing Unit (GPU) with 10GB Video Random Access

Memory (VRAM), supported by 64GB of Random Access Memory (RAM). Prior to training, audio recordings were segmented into 3-s chunks and converted into 80-dimensional Mel-filterbank features. To balance the dataset, each speaker was limited to between 8 and 100 utterances, preventing overrepresentation of speakers with extensive data. Then the augmentation techniques were applied. The model was optimized using the Adam optimizer, with classification error used as the primary training objective throughout the two-stage training process.

- Stage 1: Training began with the embedding layers frozen and a batch size of 8. The initial learning rate was set to 0.001 and linearly decreased during the first four epochs.
- Stage 2: All layers were subsequently unfrozen for extended training over 25 epochs, while a smaller learning rate was used, beginning at 0.0001 and decreasing linearly to 0.00001 to minimize disruption to the pretrained weights.

Evaluation after each epoch indicated that the best trade-off between performance and computational efficiency was achieved after 12 epochs.

IV. RESULTS AND DISCUSSION

The primary metric used was the EER, a widely accepted indicator in speaker verification tasks, defined as the error rate at which the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). During implementation, the current work used cosine similarity between length-normalized speaker embeddings following standard VoxCeleb evaluation practice [3-5] and ECAPA/x-vector scoring conventions [1, 2] to identify the threshold at which FAR and FRR intersect. The resulting EER values are reported in Tables III and IV, with lower values indicating better speaker discrimination performance. Specifically, Table III compares the EER performance of the baseline and fine-tuned models across the four evaluation datasets, where the fine-tuned model achieved consistent improvements on the Arabic datasets, although a slight increase in EER was observed on VoxCeleb1.

TABLE III. COMPARISON OF BASELINE AND FINE-TUNED MODELS' EER (%) ACROSS DATASETS

Dataset	Baseline EER (%)	Fine-tuned model EER (%)
VoxCeleb1	0.90	0.98
ArabCeleb	4.90	4.74
QASR (Test)	8.76	6.84
In-house	5.59	5.18

Figure 1 provides a visual comparison of the average EER values for both baseline and fine-tuned models under noisy conditions, where the fine-tuned model generally demonstrated improved robustness, particularly on the more challenging Arabic datasets, which is consistent with the quantitative results in Table III. These results are also presented in detail in Table IV, showcasing the performance difference based on which exact augmentation was performed. The fine-tuned model consistently outperformed the baseline under challenging noise conditions, particularly on QASR (Test) and the in-house dataset.

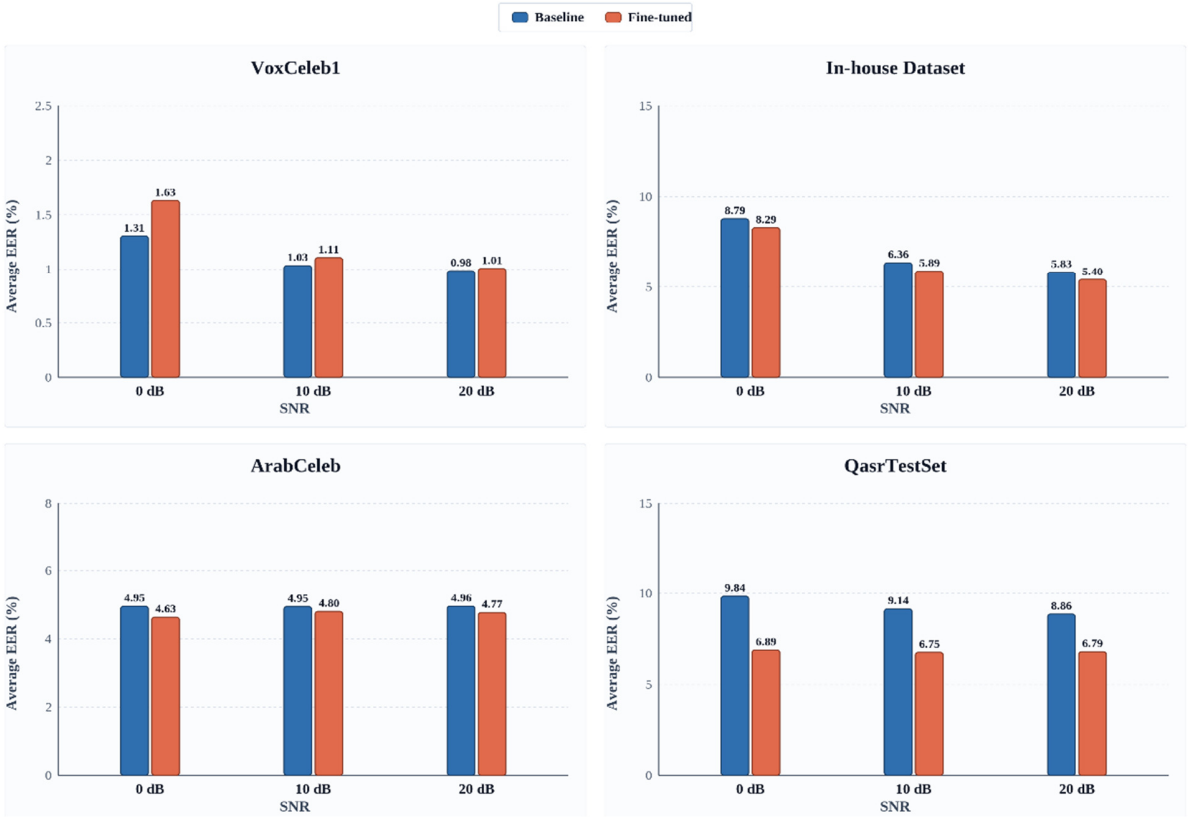

Fig. 1. Comparison of baseline vs fine-tuned models' average EER with varying added noise conditions across the test datasets.

TABLE IV. EER (%) UNDER ADDED NOISE (0, 10, 20 DB)

Dataset and Model	Noise			Music			Babble		
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB
VoxCeleb1 | Baseline	1.3	1.1	1.0	1.3	1.0	1.0	1.3	1	0.9
VoxCeleb1 | Fine-tuned	1.6	1.1	1.0	1.6	1.0	0.9	1.7	1.1	1.0
In-house | Baseline	9.3	6.3	5.7	8.3	6.4	5.9	8.7	6.5	5.9
In-house | Fine-tuned	8.7	5.8	5.3	7.5	5.8	5.4	8.6	6.0	5.5
ArabCeleb | Baseline	5.3	5.1	5.1	4.8	4.9	4.9	4.8	4.9	4.9
ArabCeleb | Fine-tuned	4.7	4.8	4.8	4.5	4.8	4.7	4.7	4.8	4.8
QASR (Test) | Baseline	10.5	9.3	9.2	9.5	8.8	8.8	9.5	9.3	8.6
QASR (Test) | Fine-tuned	7.1	6.7	6.7	6.8	6.7	6.8	6.7	6.8	6.9

Figure 2 further illustrates the reduced sensitivity of the fine-tuned model to additive noise. However, the improvements were more pronounced at lower SNR levels, whereas the performance gap decreased at higher SNR levels, suggesting that additional refinement is still required to maintain consistent robustness across all acoustic conditions.

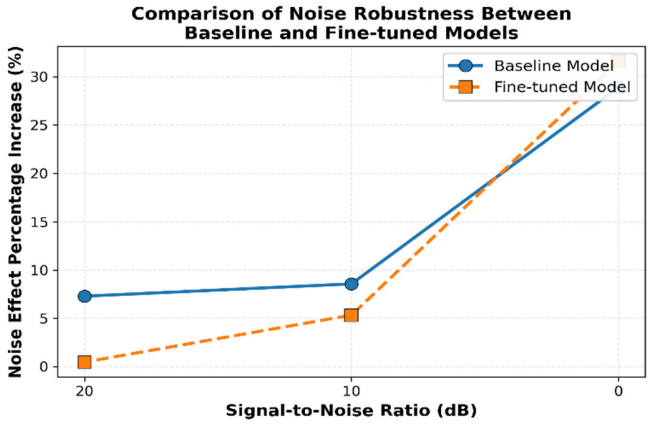

Fig. 2. Comparison of noise robustness between baseline and fine-tuned models.

Overall, the fine-tuned model demonstrated clear improvements on Arabic datasets, despite a slight reduction in performance on the English benchmark. Nevertheless, this trade-off produced a more balanced multilingual speaker verification system. The observed performance differences between Arabic and English datasets further highlight the challenges associated with multilingual speaker verification, particularly when adapting models originally trained on English speech to languages with distinct phonetic structures and dialectal diversity. Finally, the EER values reported were aggregated across all speakers without gender-based separation, although such an investigation represents an important direction for future research.

V. CONCLUSION

This paper introduced enhancements to the Emphasized Channel Attention, Propagation and Aggregation - Time Delay Neural Network (ECAPA-TDNN) model aimed at improving multilingual speaker verification, with a particular focus on Arabic. The model was trained on the Voice of Celebrities 1 (VoxCeleb1) and VoxCeleb2 corpora combined with the Arabic Qatar Computing Research Institute (QCRI) Aljazeera Speech Resource (QASR), and was evaluated on four held-out test sets: the VoxCeleb1 (Vox1-O) protocol, the Arab Celebrity (ArabCeleb) evaluation files, a held-out QASR test split with unseen speakers, and an in-house Arabic dataset of authentic recordings collected by T2 Company. The key novelty of this work lies in adapting a state-of-the-art speaker verification architecture, originally optimized for English, to the Arabic language by integrating the Arabic datasets, expanding the classifier from 7,205 to 9,707 classes, and employing a two-stage fine-tuning strategy with frozen and unfrozen embedding layers. Implementing this strategy resulted in notable improvements in Arabic speaker recognition while mitigating catastrophic forgetting of English capabilities. Specifically, the fine-tuned model reduced the Equal Error Rate (EER) on the QASR (Test) from 8.76% to 6.84%, representing a relative improvement of approximately 22%. On the ArabCeleb dataset, the EER decreased from 4.90% to 4.74%, and on the proposed in-house Arabic dataset, it improved from 5.59% to 5.18%. While a slight increase was observed on VoxCeleb1 (from 0.90% to 0.98%), this trade-off is expected given the increased emphasis on Arabic phonetics and demonstrates the inherent challenge of developing truly multilingual systems. Furthermore, the model exhibited improved robustness under noisy conditions, particularly at lower Signal-to-Noise Ratio (SNR) levels across noise, music, and babble distortions, confirming its suitability for real-world deployment scenarios.

This work addresses the gap in Arabic speaker verification by adapting the ECAPA-TDNN model to Arabic-specific datasets, improving its performance, and providing a foundation for future research in this area. It also demonstrates that high-quality results can be achieved with a cost-efficient hardware setup, proving that significant resource investments are not always required. While the improvements in Arabic and noise robustness are promising, further refinement is needed to ensure consistent performance across all languages and conditions. Future work will explore gender-specific performance analysis, integration of additional Arabic dialect-specific datasets, and advanced domain adaptation techniques to further reduce the English-Arabic performance gap.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

ACKNOWLEDGMENT

This work was funded and supported by the Research and Innovation Department at T2 Company.

DATA AVAILABILITY

The publicly available datasets used in this study can be accessed from their citations: VoxCeleb1 and VoxCeleb2 from the VGG group at the University of Oxford [3-5], ArabCeleb from its corresponding publication [17], and QASR from the Qatar Computing Research Institute [18]. The in-house Arabic dataset used in this study is proprietary to T2 Company and is not publicly available because of confidentiality and licensing restrictions. Additional information regarding the in-house dataset may be obtained from the corresponding author upon reasonable request and subject to approval by T2 Company.

REFERENCES

[1] 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)*, Calgary, AB, Canada, Apr. 2018, pp. 5329–5333, https://doi.org/10.1109/ICASSP.2018.8461375.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech 2020*, Shanghai, China, Oct. 2020, pp. 3830–3834, https://doi.org/10.21437/Interspeech.2020-2650.

[3] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech 2017*, Stockholm, Sweden, Aug. 2017, pp. 2616–2620, https://doi.org/10.21437/Interspeech.2017-950.

[4] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech 2018*, Hyderabad, India, Sept. 2018, pp. 1086–1090, https://doi.org/10.21437/Interspeech.2018-1929.

[5] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, Mar. 2020, Art. no. 101027, https://doi.org/10.1016/j.csl.2019.101027.

[6] N. R. Koluguri, T. Park, and B. Ginsburg, "TitaNet: Neural Model for Speaker Representation with 1D Depth-Wise Separable Convolutions and Global Context," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 8102–8106, https://doi.org/10.1109/ICASSP43922.2022.9746806.

[7] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking," in *Interspeech 2023*, Dublin, Ireland, Aug. 2023, pp. 5301–5305, https://doi.org/10.21437/Interspeech.2023-1513.

[8] W. Zhu *et al.*, "SpeechNAS: Towards Better Trade-Off Between Latency and Accuracy for Large-Scale Speaker Verification," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Cartagena, Colombia, Dec. 2021, pp. 1102–1109, https://doi.org/10.1109/ASRU51503.2021.9688017.

[9] J.-H. Kim, J. Heo, H. Shim, and H.-J. Yu, "Extended U-Net for Speaker Verification in Noisy Environments," in *Interspeech 2022*, Incheon, South Korea, Sept. 2022, pp. 590–594, https://doi.org/10.21437/Interspeech.2022-155.

[10] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 1021–1028, https://doi.org/10.1109/SLT.2018.8639585.

[11] W. Helali, Z. Hajaiej, and A. Cherif, "Real Time Speech Recognition based on PWP Thresholding and MFCC using SVM," *Engineering, Technology & Applied Science Research*, vol. 10, no. 5, pp. 6204–6208, Oct. 2020, https://doi.org/10.48084/etasr.3759.

[12] S. Nawaz *et al.*, "Cross-modal Speaker Verification and Recognition: A Multilingual Perspective," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, June 2021, pp. 1682–1691, https://doi.org/10.1109/CVPRW53098.2021.00184.

[13] K. Nam, Y. Kim, J. Huh, H.-S. Heo, J. Jung, and J. S. Chung, "Disentangled Representation Learning for Multilingual Speaker Recognition," in *Interspeech 2023*, Dublin, Ireland, Aug. 2023, pp. 5316–5320, https://doi.org/10.21437/Interspeech.2023-1603.

[14] H. Zhang, L. Wang, K. A. Lee, M. Liu, J. Dang, and H. Chen, "Meta-Learning for Cross-Channel Speaker Verification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, June 2021, pp. 5839–5843, https://doi.org/10.1109/ICASSP39728.2021.9413978.

[15] S. G. Kruthika, C. N. Trisiladevi, and P. Mahesha, "Voice Comparison Approaches for Forensic Application: A Review," in *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India, May 2023, pp. 797–802, https://doi.org/10.1109/ICSCCC58608.2023.10176553.

[16] A. Akram, M. Stanojevic, M. Ehghaghi, and J. Novikova, "Zero-Shot Multi-Lingual Speaker Verification in Clinical Trials." arXiv, Apr. 2024, https://doi.org/10.48550/arXiv.2404.01981.

[17] S. Bianco *et al.*, "ArabCeleb: Speaker Recognition in Arabic," in *AIxIA 2021 – Advances in Artificial Intelligence*, vol. 13196, S. Bandini, F. Gasparini, V. Mascardi, M. Palmonari, and G. Vizzari, Eds. Cham: Springer International Publishing, 2022, pp. 338–347.

[18] H. Mubarak, A. Hussein, S. A. Chowdhury, and A. Ali, "QASR: QCRI Aljazeera Speech Resource A Large Scale Annotated Arabic Speech Corpus," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2274–2285, https://doi.org/10.18653/v1/2021.acl-long.177.

[19] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus." arXiv, Oct. 2015, https://doi.org/10.48550/arXiv.1510.08484.

[20] M. Ravanelli *et al.*, "Open-Source Conversational AI with SpeechBrain 1.0," *Journal of Machine Learning Research*, vol. 25, no. 333, pp. 1–11, 2024.