An Efficient Uzbek Speaker Recognition System for Resource-Constrained Devices Using Compact Acoustic Features and Lightweight Deep Models

Parakhat Nurimov
Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, Tashkent, Uzbekistan
paranur87@gmail.com (corresponding author)

Narzillo Mamatov
Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, National Research University, Tashkent, Uzbekistan
m_narzullo@mail.ru

Received: 9 April 2026 | Revised: 28 April 2026 | Accepted: 9 May 2026

 DOI: https://doi.org/10.48084/etasr.19226

ABSTRACT

Speaker recognition systems have achieved strong performance, but many high-performing approaches remain computationally expensive and therefore not well-suited to resource-constrained devices. This limitation is particularly important in low-resource settings, including Uzbek speech applications, where practical lightweight solutions remain limited. This study presents an efficient Uzbek closed-set, text-independent speaker identification framework based on compact acoustic features and lightweight deep models. Two acoustic representations, namely MFCC-13 and Log-Mel-40, were evaluated along with two lightweight convolutional architectures, namely Small CNN and Compact CNN. The systems were assessed for recognition accuracy, F1 Score, parameter count, model size, and inference latency. The experimental results showed that the Log-Mel-40 + Compact CNN configuration achieved the best overall performance, obtaining 96.44% accuracy and 0.8957 F1-score, while maintaining a compact model size of 0.4606 MB and low inference latency. The findings indicate that practical Uzbek speaker recognition can be achieved on resource-constrained platforms through an appropriate combination of compact acoustic features and lightweight deep models.

Keywords-speaker recognition; Uzbek speech; lightweight deep learning; MFCC; Log-Mel spectrogram; resource-constrained devices; compact convolutional neural networks

I. INTRODUCTION

Speaker recognition is an important speech processing task with applications in biometric authentication, access control, forensics, and personalized digital services [1, 2]. Deep learning has substantially improved speaker recognition performance through neural embedding and convolution-based frameworks that learn speaker-discriminative information directly from acoustic inputs [1-4]. However, many high-performing systems remain computationally expensive, which limits their practical use on mobile, embedded, and other resource-constrained devices [1, 6, 7].

For lightweight deployment scenarios, model efficiency is as important as recognition quality. In such settings, the objective is not only to achieve high recognition accuracy, but also to maintain low model size and low inference latency [5-8]. Constrained convolutional architectures, lightweight speaker verification models, and quantization-based approaches can provide practical trade-offs between performance and deployment cost [5-8, 13, 14]. Another important factor is the acoustic representation. MFCCs are widely used because they are compact and computationally efficient, whereas Log-Mel representations preserve richer time-frequency information and can provide stronger discrimination when paired with suitable neural models [1, 9, 10, 12]. In practice, however, the benefit of a richer representation depends on whether the back-end model has sufficient capacity to exploit it effectively [1, 10].

Although speaker recognition has been extensively studied on widely employed benchmarks such as VoxCeleb [3, 4], lightweight Uzbek speaker recognition remains underexplored. In particular, there is limited evidence on which compact

feature-model combinations provide the best performance-efficiency trade-off for Uzbek speech under hardware constraints. This gap is important from an application perspective, because practical deployment requires a balanced solution rather than a purely high-capacity model [4, 5, 11, 13, 14]. Speaker identification and speech recognition in local language settings have also been addressed using classical approaches, further highlighting the practical relevance of speech technologies in underexplored languages [15].

To address this gap, the present study investigates a lightweight Uzbek speaker recognition framework based on two compact acoustic representations, MFCC-13 and Log-Mel-40, and two lightweight convolutional architectures, Small CNN and Compact CNN. The evaluated systems are compared in terms of recognition accuracy, F1-score, parameter count, model size, and inference latency. The objective is to identify the most practical configuration for resource-constrained Uzbek speaker recognition.

The main contributions of this work are:

- A lightweight Uzbek speaker recognition framework for resource-constrained deployment is presented.
- Two compact acoustic representations and two lightweight convolutional architectures are comparatively evaluated under the same experimental setting.
- A practical feature-model configuration is identified by jointly considering recognition performance and computational efficiency.

II. PROPOSED METHODOLOGY

This work proposes a lightweight Uzbek speaker recognition framework for resource-constrained deployment. The method combines compact acoustic representations with lightweight convolutional neural architectures to identify a configuration that provides a practical balance between recognition quality and computational efficiency. Modern deep-learning-based speaker recognition systems commonly rely on neural feature extraction and discriminative classification backends, while compact and constrained models are particularly relevant when deployment costs need to be controlled [1, 2, 5, 13, 14].

The overall framework of the proposed system is illustrated in Figure 1. The pipeline consists of four stages: audio preprocessing, acoustic feature extraction, lightweight neural modeling, and speaker classification. First, each input utterance is converted to mono, resampled to 16 kHz, and standardized to a fixed duration of 3 s using zero-padding or truncation when necessary. This fixed-length preprocessing simplifies feature extraction, guarantees consistent input dimensions, and is suitable for lightweight systems [5]. Second, one of two compact acoustic representations is extracted: MFCC-13 or Log-Mel-40. MFCC-13 provides a compact cepstral representation with low input dimensionality and is an effective baseline in speech and speaker processing [9], whereas Log-Mel-40 preserves richer time-frequency information and is more suitable for convolution-based analysis [12]. The use of both representations makes it possible to compare a more compact cepstral front-end with a richer spectrogram-based front-end under the same training and evaluation conditions [1, 12].

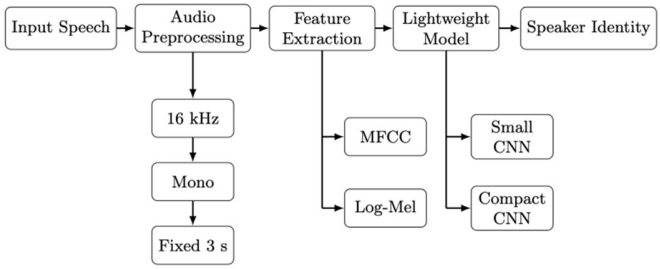

Fig. 1. Overall framework of the proposed lightweight Uzbek speaker recognition system.

To evaluate the role of model capacity under resource constraints, two lightweight convolutional architectures are considered: Small CNN and Compact CNN. The Small CNN serves as the lower-complexity baseline. It contains three convolutional stages with channel progression $1 \rightarrow 16 \rightarrow 32 \rightarrow 64$, batch normalization, ReLU activation, max-pooling in the first two stages, global adaptive average pooling, and a compact fully connected classifier with dropout. The Compact CNN is a slightly deeper and more expressive lightweight alternative. It contains five convolutional layers with channel progression $1 \rightarrow 24 \rightarrow 24 \rightarrow 48 \rightarrow 48 \rightarrow 96$, batch normalization after each convolution, ReLU activations, max-pooling after the second and fourth convolutions, adaptive average pooling, and a larger fully connected classifier with dropout. Thus, the Compact CNN is not a depthwise-separable or externally borrowed architecture, but a custom lightweight convolutional model that is deeper and more expressive than the Small CNN while remaining compact in absolute model size. This design choice is consistent with prior work on constrained and lightweight speaker recognition architectures [5, 10, 13, 14].

The task is formulated as closed-set speaker identification of specific individual speakers. In other words, the objective is to classify an input utterance into one of the predefined enrolled speaker identities, not to determine the speaker's gender. Let the number of speakers in the training set be C. Each speaker corresponds to one class. Given an input feature tensor X, the selected neural model produces a vector of logits:

$$\mathrm{z} = f(X)$$

where $f(\cdot)$ denotes the chosen lightweight neural architecture and $\mathrm{z} \in \mathbb{R}^C$. The posterior probability for the speaker class c is computed using the softmax function:

$$p(\mathrm{c}|X) = \frac{\exp(z_c)}{\sum_{k=1}^{C} \exp(z_k)}$$

The predicted speaker identity is then obtained as:

$$\hat{c} = \underset{c}{\operatorname{argmax}} p(\mathrm{c}|X)$$

Training is performed using the cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log p(\mathrm{c}|X)$$

where y_c denotes the target class label in one-hot form. This formulation follows the standard discriminative classification setting widely used in neural speaker recognition systems [1, 2, 10].

Rather than introducing a heavy or highly optimized model, the proposed methodology focuses on a controlled comparison between two acoustic representations and two lightweight neural architectures. This design makes it possible to analyze whether richer spectral input improves Uzbek speaker discrimination and whether a moderate increase in model capacity yields a more favorable performance-efficiency trade-off under strict resource constraints [5, 12-14].

III. EXPERIMENTAL SETUP

The study was designed as a controlled comparison between compact acoustic representations and lightweight neural architectures for Uzbek speaker recognition.

A. Dataset and Experimental Settings

The experiments were conducted in a closed-set speaker identification setting, where each class corresponds to a specific individual speaker. Thus, the task is to identify one speaker among a predefined set of enrolled speakers and should not be confused with speaker gender classification. The speech setting in this study is text-independent, meaning that the system is designed to recognize speaker identity without requiring a fixed or predefined sentence. The dataset was organized using an utterance-level split with a shared speaker label space, meaning that the same set of speakers was preserved across the training, validation, and test subsets, while their utterances were partitioned across these subsets.

In total, 226 speakers were included, with 22,509, 4,821, and 4,830 utterances in the training, validation, and test subsets, respectively. On average, each speaker contributed approximately 100 training utterances and 21 utterances to both the validation and test subsets. The main dataset and experimental settings are summarized in Table I. Similar speaker recognition studies commonly rely on fixed data splits and stable class definitions to ensure fair comparison across models [3, 4].

B. Input Preparation and Configurations

All audio recordings were converted to mono, resampled to 16 kHz, and normalized to a fixed duration of 3 s. Short utterances were zero-padded, and long utterances were truncated, ensuring consistent input size for all models. After preprocessing, two compact acoustic representations were extracted from each utterance: MFCC-13 and Log-Mel-40. The extracted features were stored as .npy files and indexed through metadata CSV files for the training, validation, and test subsets. This guaranteed that all compared systems used the same split and preprocessing pipeline, with only the feature representation and neural model being varied [9].

The main experiments consisted of four feature-model combinations: MFCC-13 + Small CNN, Log-Mel-40 + Small CNN, MFCC-13 + Compact CNN, and Log-Mel-40 + Compact CNN. This compact comparison design was chosen to analyze two factors systematically: the effect of the acoustic representation under a fixed architecture, and the effect of model capacity under a fixed feature type.

TABLE I. DATASET AND EXPERIMENTAL SETTINGS

Parameter	Value
Recognition setting	Closed-set speaker identification of specific individuals
Speech setting	Text-independent
Split strategy	Utterance-level split with shared speaker label space
Speakers (train/validation/test)	226 / 226 / 226
Utterances (train/validation/test)	22,509 / 4,821 / 4,830
Average samples per speaker (train/validation/test)	~100 / ~21 / ~21
Audio setup	Mono, 16 kHz, 3 s
Length normalization	Zero-padding for short utterances; truncation for long utterances
Acoustic representations	MFCC-13, Log-Mel-40
Neural architectures	Small CNN, Compact CNN
Evaluation metrics	Accuracy, F1-score, model size, and inference latency

C. Training and Evaluation Protocol

All models were trained under the same optimization protocol. The task was formulated as a standard multi-class speaker classification problem with 226 output classes, and training was performed using the cross-entropy loss function. The optimization process used the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 1×10^{-4}. The batch size was set to 32, and the maximum number of training epochs was 30. To stabilize learning, a cosine annealing learning-rate schedule was applied with $T_{max} = 30$ and $\eta_{min} = 10^{-5}$. Model selection was based on validation accuracy, and the checkpoint with the highest validation accuracy was retained for final test evaluation. Early stopping with a patience of 7 epochs was used to reduce unnecessary training and limit overfitting. The same random seed (42) and validation policy were applied to all four configurations to ensure a fair comparison. All experiments were conducted on a Mac mini M4 using the Apple Silicon MPS backend in PyTorch, and all reported latency values were measured under the same hardware and runtime environment.

D. Evaluation Metrics

The evaluated systems were assessed using both recognition-oriented and efficiency-oriented metrics. The recognition metrics were accuracy (%) and F1-score, while the efficiency metrics included parameter count, model size (MB), inference latency (ms/sample), and total inference time (s). Among these, per-sample inference latency was treated as the

most informative runtime-oriented measure. The overall objective of the experimental design was not simply to maximize recognition accuracy, but to identify the feature-model combination that provides the most favorable performance-efficiency trade-off for lightweight Uzbek speaker recognition.

IV. RESULTS AND DISCUSSION

The experimental results obtained from the four evaluated configurations are presented in Tables II and III and Figure 2. The study compared two compact acoustic representations, namely MFCC-13 and Log-Mel-40, with two lightweight convolutional architectures, namely Small CNN and Compact CNN. The objective was to identify the most suitable configuration for Uzbek speaker recognition under resource-constrained conditions by jointly considering recognition quality and computational efficiency.

A. Experimental Results

Table II presents the main recognition results of the evaluated feature-model combinations. The results show a difference between the two neural architectures. In both feature settings, the Compact CNN outperformed the Small CNN by a considerable margin. The best-performing configuration was Log-Mel-40 + Compact CNN, which achieved 96.44% accuracy and 0.8957 F1-score. The second-best configuration was MFCC-13 + Compact CNN, which achieved 94.78% accuracy and 0.8536 F1-score. In contrast, the Small CNN models had noticeably lower performance, with 90.00% and 90.31% accuracy for MFCC-13 and Log-Mel-40, respectively.

TABLE II. PERFORMANCE COMPARISON OF THE EVALUATED CONFIGURATIONS FOR UZBEK SPEAKER RECOGNITION

Exp.	Feature	Model	Accuracy (%)	F1-score	Parameter	Model size (MB)
E1	MFCC-13	Small CNN	90.00	0.7478	42,370	0.1625
E2	Log-Mel-40	Small CNN	90.31	0.7328	42,370	0.1625
E3	MFCC-13	Compact CNN	94.78	0.8536	120,266	0.4606
E4	Log-Mel-40	Compact CNN	96.44	0.8957	120,266	0.4606

TABLE III. EFFICIENCY ANALYSIS OF THE EVALUATED CONFIGURATIONS IN TERMS OF INFERENCE LATENCY

Exp.	Feature	Model	Latency (ms/sample)	Inference time (s)
E1	MFCC-13	Small CNN	0.0158	1.2538
E2	Log-Mel-40	Small CNN	0.0165	2.1641
E3	MFCC-13	Compact CNN	0.0406	1.7902
E4	Log-Mel-40	Compact CNN	0.0239	3.7396

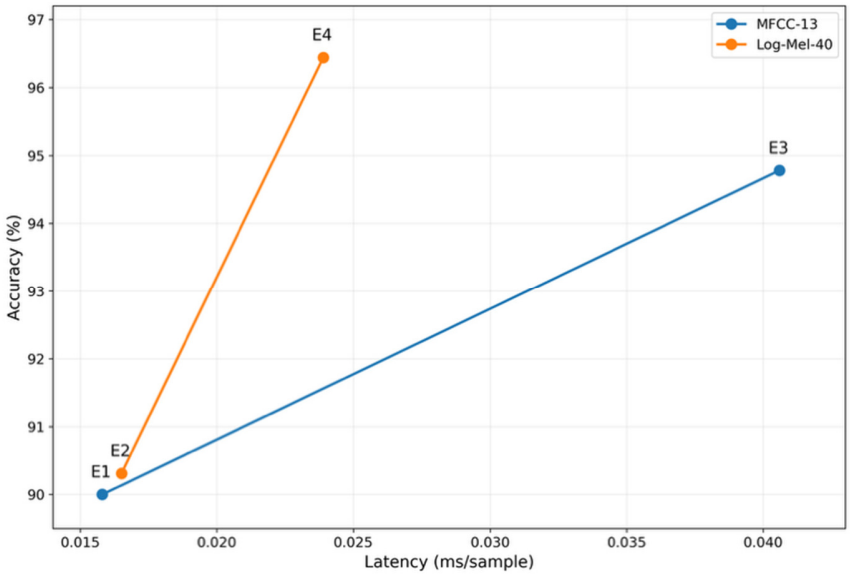

Fig. 2. Accuracy-latency trade-off of the evaluated configurations.

These results indicate that the evaluated Uzbek speaker identification task benefited strongly from a slightly more expressive lightweight architecture. Although the Small CNN provided a valid baseline, the Compact CNN offered a substantial improvement in recognition quality while still preserving a very small model footprint. The results, therefore, confirm that lightweight speaker recognition can achieve strong performance without relying on computationally expensive architectures.

B. Effect of Acoustic Features

The comparison between MFCC-13 and Log-Mel-40 reveals that the usefulness of the acoustic representation depended on the model architecture. In the Small CNN setting,

the difference between the two feature types was limited. Log-Mel-40 slightly improved the accuracy from 90.00% to 90.31%, but the corresponding F1-score was slightly lower than that of MFCC-13. This suggests that in a highly constrained baseline network, the richer spectral structure of Log-Mel-40 could not be fully exploited.

A different pattern emerged when the Compact CNN was used. In that setting, Log-Mel-40 outperformed MFCC-13 in both accuracy and F1-score. The accuracy increased from 94.78% to 96.44%, while the F1-score improved from 0.8536 to 0.8957. This behavior indicates that the richer time-frequency information contained in the Log-Mel representation becomes more beneficial when paired with a model that has sufficient representational capacity.

From these observations, it can be concluded that Log-Mel-40 is the more effective acoustic representation for the considered Uzbek speaker recognition task when combined with a sufficiently expressive lightweight model, whereas its advantage is less evident in a simpler baseline architecture. This observation is consistent with prior work, which suggests that richer acoustic representations become more useful when paired with more expressive speaker embedding back-ends [10, 12].

C. Effect of Model Architecture

The influence of the model architecture was stronger than the influence of feature selection alone. Replacing the Small CNN with the Compact CNN yielded large improvements for both feature types. For MFCC-13, the accuracy increased from 90.00% to 94.78%, corresponding to an absolute gain of 4.78%. For Log-Mel-40, the same architectural replacement increased the accuracy from 90.31% to 96.44%, corresponding to an absolute gain of 6.13%. A similar trend can be seen in the F1-score values, which also increased substantially when the Compact CNN was used.

These gains indicate that the Compact CNN can better model speaker-discriminative patterns than the Small CNN, despite remaining compact in absolute terms. This is an important finding for the targeted deployment scenario. It suggests that a moderate increase in model complexity can provide a significant increase in speaker recognition performance while keeping the resulting model suitable for low-resource environments. The observed gain is also in line with prior lightweight speaker verification studies, where carefully designed compact architectures have been shown to provide stronger performance than simpler small-footprint baselines [5, 13, 14].

D. Efficiency Analysis

The efficiency-related results are outlined in Table III. The Small CNN models had the lowest per-sample inference latency, achieving 0.0158 ms/sample for MFCC-13 and 0.0165 ms/sample for Log-Mel-40. The Compact CNN configurations required more computation, with 0.0406 ms/sample for MFCC-13 and 0.0239 ms/sample for Log-Mel-40. However, these latency values remained very low in absolute terms. The total inference time values varied among the configurations. This is expected, since total runtime depends not only on the model itself, but also on implementation details, execution order, and runtime environment. For that reason, per-sample latency is considered the more informative efficiency indicator in this study.

Although the Compact CNN models were slower than the Small CNN baselines, their model size remained only 0.4606 MB, which is still highly compact and appropriate for deployment in resource-constrained settings. Therefore, the efficiency analysis suggests that the computational overhead introduced by the Compact CNN is modest relative to the substantial gain in recognition performance. From a broader lightweight speaker verification perspective, the reported model size of 0.4606 MB remains competitive for deployment-oriented scenarios, where compact architectures and quantization have been emphasized as practical strategies [6, 7, 13, 14].

E. Performance-Efficiency Trade-Off

A direct comparison of recognition quality and efficiency is illustrated in Figure 2, which plots recognition accuracy against inference latency. It is observed that the Small CNN models occupy the low-latency region but remain limited in recognition performance, with accuracy values close to 90%. In contrast, the Compact CNN models shift toward a more favorable accuracy region, particularly when paired with Log-Mel-40.

Taken together, the results indicate that Log-Mel-40 + Compact CNN provides the most favorable balance between recognition quality and computational efficiency among the evaluated configurations. MFCC-13 + Compact CNN can be regarded as the second-best alternative, while the two Small CNN models remain more limited in overall recognition performance despite their lower latency. This finding is consistent with recent compact speaker verification studies, which suggest that the best deployment-oriented solution is often not the smallest model, but rather the one that provides the most favorable balance between recognition quality, memory footprint, and computational cost [13, 14].

These results indicate that the smallest and fastest model is not necessarily the most practical one when recognition quality is taken into account. In the present study, the best overall solution was achieved by a slightly more complex but still lightweight configuration.

F. Best Configuration

Based on the experimental results, the best configuration for the considered task was Log-Mel-40 + Compact CNN. This model achieved 96.44% accuracy, 0.8957 F1-score, 0.0239 ms/sample inference latency, and a model size of only 0.4606 MB. These results demonstrate that the proposed configuration provides the most favorable balance between recognition quality and computational efficiency.

The MFCC-13 + Compact CNN configuration may still be thought of as a valid alternative, especially when a simpler feature representation is preferred. However, the results indicate that the Log-Mel-based system is superior when paired with the more expressive lightweight architecture. From a practical perspective, the best-performing configuration

remains compact enough to be considered for deployment on low-resource devices, while providing substantially higher recognition performance than the simpler baselines.

G. Discussion

Several important observations can be drawn from the obtained results. First, the study demonstrates that high Uzbek speaker recognition performance can be achieved using lightweight models. The best-performing model remained below 0.5 MB, which is highly favorable for real-world deployment on devices with limited memory and computational resources. Second, the results show that feature richness alone is not sufficient. The advantage of Log-Mel-40 became visible only when it was paired with the Compact CNN. This suggests that acoustic representation and model capacity must be considered jointly rather than independently.

Third, the comparison highlights that the lowest latency does not automatically correspond to the best practical solution. Although the Small CNN achieved the fastest inference, its recognition performance was lower. In contrast, the Compact CNN introduced only a moderate increase in computational cost while providing a substantial gain in accuracy and F1-score. Finally, the findings confirm that the combination of compact acoustic features and lightweight deep models can offer an effective and practical solution for Uzbek speaker recognition under resource constraints. In particular, the Log-Mel-40 + Compact CNN configuration can be regarded as the most suitable solution among the evaluated alternatives.

Despite the encouraging results, this study has several limitations. First, the experimental comparison was restricted to four lightweight feature-model configurations and did not include external baselines such as TDNN-based or larger embedding-based speaker recognition systems. Therefore, the reported findings should be interpreted as a controlled comparison within a lightweight Uzbek speech setting rather than as a claim of superiority over the broader speaker recognition literature or as a large-scale state-of-the-art benchmarking study. Second, the reported results are based on a single fixed experimental split and one training run per configuration. Although the same protocol was consistently applied to all compared models, additional repeated runs and broader evaluation settings would provide a more comprehensive estimate of robustness and result stability. Third, the present study does not include validation on public benchmark datasets or robustness evaluation under noisy conditions. In addition, the feature space was intentionally restricted to two compact acoustic representations to maintain a focused, lightweight comparison. These aspects are important for broader generalization and remain relevant directions for future research.

V. CONCLUSION

This study investigated a lightweight Uzbek speaker recognition framework based on compact acoustic features and lightweight convolutional neural models. Two acoustic representations, MFCC-13 and Log-Mel-40, together with two lightweight architectures, Small CNN and Compact CNN, were comparatively evaluated under the same closed-set, text-independent speaker identification protocol.

The experimental results showed that the Log-Mel-40 + Compact CNN configuration achieved the best overall performance, reaching 96.44% accuracy and 0.8957 F1-score, while maintaining a compact model size of 0.4606 MB and low inference latency. The results also showed that model capacity had a strong influence on recognition performance, and that the richer Log-Mel representation became more effective when paired with the more expressive lightweight architecture.

From a broader practical perspective, the study confirms that lightweight Uzbek speaker recognition does not necessarily require large or computationally expensive models. Instead, a carefully selected combination of compact acoustic features and lightweight deep models can provide a favorable balance between recognition quality, memory footprint, and runtime efficiency. Among the evaluated configurations, Log-Mel-40 + Compact CNN provided the most favorable performance-efficiency trade-off and can be considered the most suitable solution for lightweight deployment.

These findings are especially relevant for resource-constrained and on-device speech applications, where both recognition quality and computational efficiency must be considered jointly. At the same time, the present results should be interpreted within the scope of a controlled lightweight comparison on Uzbek speech, without public-benchmark validation or noisy-condition evaluation. For future work, the framework should be extended with broader Uzbek speech conditions, additional efficient architectures, repeated runs for stronger robustness analysis, validation on public benchmark datasets, and more deployment-oriented evaluations on embedded or mobile platforms.

DECLARATION OF COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENT

Not applicable to this work.

DATA AVAILABILITY

The dataset used in this study is available from the corresponding author upon reasonable request.

AI USE AND DECLARATION OF GENERATIVE AI USE

During the preparation of this work, generative AI tools were used for language refinement and grammatical editing. After using these tools, the authors reviewed and edited the manuscript as needed and take full responsibility for the content of the publication.

REFERENCES

[1] Z. Bai and X.-L. Zhang, "Speaker Recognition Based on Deep Learning: An Overview," *Neural Networks*, vol. 140, pp. 65–99, Aug. 2021, https://doi.org/10.1016/j.neunet.2021.03.004.

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

[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] 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.

[5] N. Simić *et al.*, "Speaker Recognition Using Constrained Convolutional Neural Networks in Emotional Speech," *Entropy*, vol. 24, no. 3, Mar. 2022, Art. no. 414, https://doi.org/10.3390/e24030414.

[6] B. Liu, H. Wang, and Y. Qian, "Towards Lightweight Speaker Verification via Adaptive Neural Network Quantization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3771–3784, 2024, https://doi.org/10.1109/TASLP.2024.3437237.

[7] B. Liu, H. Wang, and Y. Qian, "Extremely Low Bit Quantization for Mobile Speaker Verification Systems Under 1MB Memory," in *Interspeech 2023*, Dublin, Ireland, Aug. 2023, pp. 1973–1977, https://doi.org/10.21437/Interspeech.2023-800.

[8] Z. Özcan and T. Kayıkçıoğlu, "Evaluating MFCC-Based Speaker Identification Systems with Data Envelopment Analysis," *Expert Systems with Applications*, vol. 168, Apr. 2021, Art. no. 114448, https://doi.org/10.1016/j.eswa.2020.114448.

[9] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980, https://doi.org/10.1109/TASSP.1980.1163420.

[10] 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.

[11] A. Amraoui and S. Saadi, "A Novel Approach on Speaker Gender Identification and Verification Using DWT First Level Energy and Zero Crossing," *Engineering, Technology & Applied Science Research*, vol. 12, no. 6, pp. 9570–9578, Dec. 2022, https://doi.org/10.48084/etasr.5269.

[12] I. McLoughlin *et al.*, "Spectrogram Features for Audio and Speech Analysis," *Applied Sciences*, vol. 16, no. 2, Jan. 2026, Art. no. 572, https://doi.org/10.3390/app16020572.

[13] M. Kim *et al.*, "Light-Weight Speaker Verification with Global Context Information," in *Interspeech 2022*, Incheon, South Korea, Sep. 2022, pp. 5105–5109, https://doi.org/10.21437/Interspeech.2022-10932.

[14] J.-H. Choi, J.-Y. Yang, and J.-H. Chang, "Efficient Lightweight Speaker Verification with Broadcasting CNN-Transformer and Knowledge Distillation Training of Self-Attention Maps," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4580–4595, 2024, https://doi.org/10.1109/TASLP.2024.3463491.

[15] N. Mamatov, A. Samijonov, P. Nurimov, and N. Niyozmatova, "Automatic Speaker Identification by Voice Based on Vector Quantization Method," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 10, pp. 2443–2445, Aug. 2019, https://doi.org/10.35940/ijitee.J9523.0881019.