

Crime-Related Keyword Spotting in Low-Resource Nigerian Languages Using Knowledge-Distilled DSCNN for Security Monitoring

ONI Oluwabunmi Ayankemi¹, AYENI Joshua Ayobami², IPEAYEDA Funmilola W.²,
OYEDIRAN Mayowa O.²

¹Department of Computer Science, The Polytechnic, Ibadan, Nigeria

²Faculty of Computing, Ajayi Crowther University, Oyo, Nigeria

DOI: <https://doi.org/10.51584/IJRIAS.2026.11080138>

Received: 02 September 2026; Accepted: 07 September 2026; Published: 18 September 2026

ABSTRACT

Insecurity remains a major challenge in Nigeria, creating a need for intelligent tools that can automatically detect crime-related keywords in recorded meetings and conversations. This paper presents a knowledge-distilled Depthwise Separable Convolutional Neural Network (D-DSCNN) system for keyword spotting in the three predominant Nigerian languages — Hausa, Igbo, and Yoruba. A custom dataset of crime-related keywords was collected from native speakers and expanded through data augmentation. The proposed system achieves 96.40% test accuracy with extremely low false positive and false negative rates, while maintaining a small computational footprint suitable for edge devices. The results demonstrate the practical feasibility of deploying efficient multilingual KWS systems for security monitoring in low-resource African language settings.

Keywords: Keyword spotting, Nigerian languages, Hausa, Igbo, Yoruba, Knowledge distillation, Security monitoring, Low-resource speech, Edge devices

INTRODUCTION

Insecurity remains one of the most pressing challenges confronting Nigeria. The country has experienced a rise in security threats linked to groups such as Boko Haram and other extremist organisations, as well as widespread incidents of kidnapping, armed robbery, and communal violence. These developments have significant implications for industrialisation, sustainable development, and social stability. In this context, the ability to automatically analyse large volumes of spoken content particularly recorded meetings, community discussions, and intercepted communications has become increasingly important for security monitoring and situational awareness.

Keyword spotting (KWS) offers a practical solution for this task. Unlike full automatic speech recognition, which attempts to transcribe entire utterances, KWS focuses on the reliable detection of a predefined set of keywords within continuous audio streams. This focused objective enables systems to operate with substantially lower computational requirements, making them suitable for deployment on resource-constrained devices commonly available in low-resource settings (López-Espejo et al., 2021; Zhang et al., 2017).

Despite the rapid progress of deep learning-based KWS systems, most existing research has concentrated on high-resource languages such as English and Chinese. African languages, including the three major Nigerian languages Hausa, Igbo, and Yoruba remain severely under-resourced in terms of annotated speech data (Inuwa-Dutse, 2025). These languages are tonal and exhibit substantial phonetic diversity, further complicating the development of accurate keyword spotting models from limited data.

This paper addresses the gap by presenting a practical multilingual keyword spotting system tailored to crime-related speech in Hausa, Igbo, and Yoruba. The system employs a response-based knowledge distillation framework in which a high-capacity Residual Network (ResNet) teacher transfers discriminative knowledge to a compact Depthwise Separable Convolutional Neural Network (DSCNN) student. The resulting Distilled DSCNN (D-DSCNN) achieves high detection accuracy while retaining a small memory footprint and low computational cost, making it suitable for both offline analysis of recorded meetings and continuous online monitoring on edge devices.

The main contributions of this work are as follows:

- We present the first knowledge-distilled keyword spotting system specifically designed for crime-related keywords in the three predominant Nigerian languages.
- We collect and describe a custom multilingual dataset of security-relevant keywords from native speakers of Hausa, Igbo, and Yoruba.
- We demonstrate that the distilled model achieves 96.40% test accuracy with extremely low false-alarm and missed-detection rates while remaining suitable for edge deployment.
- We discuss the practical implications of the system for offline and online security monitoring applications in the Nigerian context.

Related Work

Keyword spotting has been applied to a variety of security-related tasks, including the detection of prohibited words in call-centre recordings and the monitoring of sensitive speech content (Kavya and Venna, 2014; Lepak et al., 2021). Most existing systems, however, have been developed for high-resource languages and assume the availability of large annotated corpora.

Research on speech technology for African languages remains limited. Recent surveys highlight the scarcity of annotated resources for Nigerian languages and the consequent challenges for developing robust speech applications (Inuwa-Dutse, 2025). Multilingual and cross-lingual approaches that leverage pretrained speech representations have shown promise for low-resource settings (Xie et al., 2023), yet few studies have specifically targeted keyword spotting for security applications in African languages.

On the technical side, Depthwise Separable Convolutional Neural Networks have become a standard backbone for small-footprint KWS because of their favourable accuracy–efficiency trade-off (Zhang et al., 2017; Sørensen et al., 2020). Knowledge distillation has further been shown to improve the generalisation of compact models when labelled data are scarce (Tan et al., 2021; Lei et al., 2022).

The present work builds on these advances by combining response-based knowledge distillation with a DSCNN architecture and applying the resulting system to the under-explored domain of crime-related keyword spotting in Hausa, Igbo, and Yoruba.

System Design

Dataset Collection Process

Due to the absence of publicly available keyword spotting datasets for crime-related terms in Nigerian languages, a custom speech corpus was collected. Recordings were obtained from native speakers of Hausa, Igbo, and Yoruba through voluntary participation. Speakers were drawn from diverse demographic backgrounds, including variations in age, gender, and regional accents, to improve the representativeness of the data.

Participants were not required to record every keyword, resulting in a naturally unbalanced but realistic distribution of samples. In total, 1,453 original utterances were collected: 493 in Hausa, 422 in Igbo, and 538 in Yoruba. All recordings were made in controlled acoustic environments using consumer-grade microphones to minimise background noise while preserving natural speech characteristics. Audio files were stored in 16 kHz mono WAV format.

Ethical considerations were strictly observed. Informed consent was obtained from all participants prior to recording. No personal identifying information was retained, and speakers retained the right to withdraw their contributions at any time. The dataset was collected solely for research purposes related to security-oriented keyword spotting.

Due to the fact that the original dataset was relatively small, stochastic data augmentation was applied. Techniques included additive Gaussian noise and pitch shifting. Each original sample was used to generate nine augmented variants, expanding the effective training set approximately tenfold and improving model robustness to acoustic variability.

Selected Crime-Related Keywords

Ten semantically distinct keywords were selected for each of the three languages. The selection prioritised phonetic diversity, sociolinguistic relevance, and association with security-related contexts such as violence, extremism, and criminal activity. Example keywords include terms such as *ipakupa*, *igbesan*, *itajesile*, and *iponju* (Yoruba); *dumbu*, *zubardajini*, *shahada*, and *jihad* (Hausa); together with a corresponding set of ten phonetically diverse crime-related terms in Igbo. These keywords were chosen to reflect vocabulary that could appear in conversations or recordings of potential security interest, while remaining distinct enough for reliable acoustic discrimination.

Overall System Pipeline

The complete keyword spotting pipeline consists of three main stages. Firstly, feature extraction here raw audio is pre-emphasised, framed, and transformed using the Short-Time Fourier Transform. Mel-filterbank energies are computed and converted into 40-dimensional Mel-Frequency Cepstral Coefficients (MFCCs). The resulting feature matrices serve as input to the neural network. The second stage is classification with Distilled DSCNN (D-DSCNN). The core acoustic model is a Depthwise Separable Convolutional Neural Network trained via response-based knowledge distillation.

A high-capacity ResNet teacher is first trained to convergence. Its softened output distributions are then used, together with hard labels, to supervise the compact DSCNN student. At inference time, only the lightweight student model is deployed. The last stage is keyword detection where the model outputs posterior probabilities over the keyword classes. A detection decision is made by comparing the highest posterior against a predefined confidence threshold (0.70 in the present experiments). This stage supports both isolated keyword classification and continuous stream processing.

Suitability for Offline and Online Security Monitoring

The designed system is well suited to both offline and online security applications. In offline mode, the model can process pre-recorded meeting audio or call-centre recordings to flag segments containing crime-related keywords, enabling efficient review of large volumes of speech data. In online (streaming) mode, the low memory footprint (approximately 132 KB) and low computational cost (1,728 MAC operations per inference) of the D-DSCNN make continuous monitoring on edge devices feasible.

These characteristics are particularly relevant for security monitoring scenarios in Nigeria, where computational resources may be limited and real-time or near-real-time analysis of spoken content can support situational awareness. The combination of high detection accuracy and strict resource efficiency positions the system as a practical tool for assisting human analysts in identifying potentially sensitive speech content.

Experimental Evaluation

Performance of D-DSCNN on the Three Languages

The proposed Distilled Depthwise Separable Convolutional Neural Network (D-DSCNN) was evaluated on the multilingual crime-related keyword dataset covering Hausa, Igbo, and Yoruba. After data augmentation, the experimental set comprised 2,500 samples, partitioned into training (1,600), development (400), and test (500) subsets. On the held-out test set, the D-DSCNN achieved an overall accuracy of 96.40%, with precision, recall, and F1-score all equal to 0.964. These results indicate strong and balanced performance across the three languages. Although detailed per-language scores are not reported separately, the high aggregate metrics and the near-ideal diagonal dominance observed in the confusion matrix suggest that the model generalises effectively to the phonetic and tonal characteristics of Hausa, Igbo, and Yoruba. The model maintains a compact footprint of approximately 132 KB and requires only 1,728 multiply-accumulate (MAC) operations per inference, confirming its suitability for resource-constrained environments.

Class-wise Analysis and Error Patterns

Confusion-matrix analysis reveals that the D-DSCNN exhibits strong class separability. Most keyword classes are recognised with near-perfect accuracy, and off-diagonal errors are sparse and isolated. Occasional confusions occur only between a small number of acoustically similar pairs (for example, limited interchange between certain Yoruba terms such as itajesile and rogbodyan). No systematic confusion patterns across multiple classes were observed. This behaviour contrasts sharply with the baseline DSCNN, which displayed widespread misclassifications, and is superior even to the deeper ResNet teacher, which still exhibited residual errors among short-duration or phonetically close keywords.

Comparison with Baseline DSCNN and ResNet

Table 1 summarises the comparative performance of the three models evaluated under identical data splits and experimental conditions.

Table 1. Classification performance comparison

Model	Test Accuracy	Precision	Recall	F1-score	Memory	MAC Ops
DSCNN	55.20%	0.552	0.552	0.552	131.80 KB	1,728
ResNet	82.40%	0.824	0.824	0.824	1,955.79 KB	2,016
D-DSCNN	96.40%	0.964	0.964	0.964	131.79 KB	1,728

The baseline DSCNN suffers from severe overfitting, achieving high training accuracy but collapsing on the test set. The ResNet teacher improves generalisation substantially (82.40% test accuracy) but at the cost of a much larger memory footprint. The proposed D-DSCNN outperforms both models, delivering a 41.2 percentage-point absolute improvement over the baseline DSCNN and a 14.0 percentage-point improvement over the ResNet teacher, while retaining the identical computational profile of the lightweight baseline.

Practical Implications: False Alarms versus Missed Detections

In security monitoring applications, the relative costs of false positives (false alarms) and false negatives (missed detections) are critical. Table 2 reports the corresponding error rates.

Table 2. False-positive and false-negative rates

Model	False Positive Rate (FPR)	False Negative Rate (FNR)
DSCNN	4.98%	44.80%
ResNet	1.96%	17.60%
D-DSCNN	0.40%	3.60%

The D-DSCNN reduces the false-negative rate by more than 90% relative to the baseline DSCNN and achieves an extremely low false-positive rate of 0.40%. This combination is highly desirable for practical deployment: missed detections of crime-related keywords are minimised, while the system generates very few spurious alerts that would otherwise burden human analysts. Collectively, the experimental results demonstrate that the knowledge-distilled model provides both high detection reliability and the computational efficiency required for offline batch processing of recorded meetings as well as continuous online monitoring on edge devices.

DISCUSSION

Deployment Readiness for Edge Devices

The experimental results demonstrate that the proposed D-DSCNN is well positioned for deployment on resource-constrained edge devices. With a memory footprint of approximately 132 KB and only 1,728 multiply-accumulate operations per inference, the model matches the computational profile of the baseline DSCNN while delivering substantially higher accuracy (96.40% versus 55.20%). This combination of efficiency and performance is critical for always-on keyword spotting systems that must operate continuously under strict power and memory budgets.

Because the student model does not require the ResNet teacher at inference time, the final deployed system remains extremely lightweight. The low false-positive rate (0.40%) further supports practical deployment: the system is unlikely to generate excessive false alarms that would drain battery life or overwhelm downstream processing stages. These characteristics make the D-DSCNN suitable for integration into embedded platforms such as microcontrollers, smartphones, or dedicated security monitoring hardware commonly used in low-resource settings.

LIMITATIONS

Despite the strong results, several limitations should be acknowledged. First, the original dataset is relatively small (1,453 utterances before augmentation). Although extensive data augmentation was applied and the distilled model generalised well, the limited number of unique speakers and recording conditions may restrict performance under more diverse real-world acoustic environments.

Second, the current evaluation was conducted primarily under relatively clean recording conditions. Robustness to challenging noise environments—such as overlapping speech, background music, traffic, or reverberant indoor spaces—has not yet been systematically assessed. In security monitoring scenarios, audio is often captured under uncontrolled conditions, and performance degradation under noise remains an open question.

Third, the present study treats the three languages in a combined multilingual setting. Fine-grained analysis of language-specific error patterns and cross-lingual transfer effects would provide additional insights but was beyond the scope of the current experiments.

Potential Real-World Use Cases in Nigeria

The developed system has several promising applications within the Nigerian security context. One immediate use case is the offline analysis of recorded meetings, community discussions, or intercepted communications. By automatically flagging segments that contain crime-related keywords, the system can significantly reduce the time required for human analysts to review large volumes of audio.

A second application lies in continuous monitoring of public or semi-public spaces where voice activity is present, such as transportation hubs or community centres, provided appropriate legal and ethical frameworks are observed. The low computational requirements of the D-DSCNN make on-device or near-edge processing feasible, reducing the need to transmit raw audio to central servers and thereby addressing both bandwidth and privacy concerns.

Finally, the multilingual capability of the system is particularly relevant in Nigeria, where Hausa, Igbo, and Yoruba are widely spoken across different regions. A single compact model that operates effectively across these languages can support nationwide security monitoring efforts more efficiently than separate monolingual systems.

In summary, while further validation under noisy and more diverse conditions is required, the combination of high detection accuracy, extremely low false-alarm rates, and strict resource efficiency positions the proposed approach as a practical foundation for intelligent speech-based security tools in low-resource African settings.

CONCLUSION

This paper has presented a practical multilingual keyword spotting system for crime-related speech in the three predominant Nigerian languages—Hausa, Igbo, and Yoruba. By combining a carefully collected custom dataset with a response-based knowledge distillation framework, the resulting Distilled DSCNN achieves 96.40% test accuracy while maintaining a memory footprint of only approximately 132 KB and a computational cost of 1,728 MAC operations.

The system is suitable for both offline analysis of recorded meetings and continuous online monitoring on edge devices. Its low false-positive and false-negative rates make it particularly attractive for security applications, where missed detections and spurious alerts both carry significant costs. Future work will focus on expanding the dataset, evaluating robustness under realistic noise conditions, and exploring continual learning strategies that allow the system to adapt to new speakers and keywords without full retraining.

REFERENCES

1. Inuwa-Dutse, I. (2025). NaijaNLP: A survey of Nigerian low-resource languages. arXiv preprint arXiv:2502.19784.
2. Kavya, H. P., and Venna, K. (2014). Sensitive keyword spotting for crime analysis. In Proceedings of the IEEE National Conference on Communication, Signal Processing and Networking (NCCSN), pp. 1–6.
3. Lei, L., Yuan, G., Zhang, T., and Yu, H. (2022). Low-power feature-attention Chinese keyword spotting framework with distillation learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2), Article 56.
4. Lepak, Ł., Radzikowski, K., Nowak, R., and Piczak, K. J. (2021). Generalisation gap of keyword spotters in a cross-speaker low-resource scenario. *Sensors*, 21(24), 8313.
5. López-Espejo, I., Tan, Z.-H., Hansen, J. H. L., and Jensen, J. (2021). Deep spoken keyword spotting: An overview. *IEEE Access*, 10, 4169–4199.
6. Sørensen, P. M., Epp, B., and May, T. (2020). A depthwise separable convolutional neural network for keyword spotting on an embedded system. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1), Article 10.



7. Tan, Z.-H., Wang, J., Zhang, D., and Jensen, S. H. (2021). Knowledge distillation for small-footprint keyword spotting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2730–2741.
8. Xie, J., Zhang, S., Hu, Y., and Wu, X. (2023). Multilingual and cross-lingual keyword spotting with pretrained speech models. *arXiv preprint arXiv:2304.09585*.
9. Zhang, Y., Suda, N., Lai, L., and Chandra, V. (2017). Hello Edge: Keyword spotting on microcontrollers. *arXiv preprint arXiv:1711.07128*.