Response-Based Knowledge Distillation for Efficient Multilingual Keyword Spotting: A Distilled Depth wise Separable CNN Approach

Oni, Oluwabunmi Ayankemi[1], Ayeni, Joshua Ayobami[2], Makinde, Oladayo Ezekiel[3], Ojo, Olufemi Samuel[4]

[1] Department of Computer Science, the Polytechnic, Ibadan, Nigeria
[2, 3,4] Faculty of Computing, Ajayi Crowther University, Oyo, Nigeria

ABSTRACT

Keyword spotting (KWS) on resource-constrained devices remains challenging, especially for low-resource languages with limited labelled data. This paper proposes a response-based knowledge distillation framework that transfers discriminative knowledge from a high-capacity Res Net teacher to a lightweight Depth wise Separable Convolutional Neural Network (DSCNN) student. Using Mel-Frequency Cepstral Coefficients and a hybrid loss combining cross-entropy and temperature-scaled Kullback–Leibler divergence, the distilled model (D-DSCNN) achieves 96.40% test accuracy while retaining the same memory footprint (≈132 KB) and computational cost (1,728 MACs) as the baseline DSCNN. Experiments on a multilingual crime-related keyword dataset demonstrate that knowledge distillation significantly improves generalization without increasing model complexity, making the approach suitable for always-on edge deployment.

KEYWORDS

Keyword spotting, Knowledge distillation, DSCNN, Low-resource languages, Edge devices, Multilingual

I. INTRODUCTION

Keyword spotting (KWS) has become a critical enabling technology for voice-driven human–computer interaction, particularly on resource-constrained edge devices such as smartphones, smart sensors, and always-on listening systems. Unlike full automatic speech recognition (ASR), which transcribes entire utterances, KWS focuses on the reliable detection of a small set of predefined keywords or phrases within continuous audio streams. This focused objective allows KWS systems to operate with substantially lower computational and memory requirements, making them suitable for real-time, low-power deployment.

In recent years, deep neural networks—especially convolutional architectures—have significantly advanced KWS performance. Among these, the Depth wise Separable Convolutional Neural Network (DSCNN) has emerged as a particularly attractive solution for embedded applications. By factorizing standard convolutions into depth wise and point wise

operations, DSCNN drastically reduces the number of parameters and multiply–accumulate (MAC) operations while retaining competitive classification accuracy. Consequently, DSCNN-based models have been widely adopted for small-footprint keyword spotting.

Despite these advances, two interrelated challenges continue to limit the practical effectiveness of DSCNN models. First, when trained on limited labelled data, compact architectures often over fit and fail to learn robust decision boundaries. Second, the scarcity of annotated speech resources is especially acute for low-resource languages, including the major Nigerian languages Hausa, Igbo, and Yoruba. These languages are tonal and exhibit substantial phonetic diversity, further complicating the development of high-performance KWS systems from small datasets.

Knowledge distillation (KD) offers a promising strategy to mitigate these difficulties (Hinton et al., 2015). In the response-based KD paradigm, a high-capacity "teacher" network is first trained to convergence; its softened output probability distributions are then used, together with ground-truth labels, to supervise a compact "student" network. By transferring the teacher's learned inter-class relationships, the student can achieve improved generalization without any increase in model size or inference cost.

Although knowledge distillation has been successfully applied to various speech tasks, its systematic integration with DSCNN architectures for multilingual, low-resource keyword spotting remains underexplored. In particular, little attention has been paid to the joint problem of (i) designing an efficient student model suitable for edge deployment and (ii) leveraging a deeper residual teacher to overcome data scarcity in African tonal languages.

This paper addresses the above gap by proposing a response-based knowledge distillation framework in which a Residual Network (ResNet) serves as the teacher and a carefully optimized DSCNN serves as the student. The student is trained with a hybrid objective that combines conventional cross-entropy loss on hard labels with a temperature-scaled Kullback–Leibler divergence loss that aligns the student's predictions with the teacher's softened outputs. The resulting Distilled DSCNN (D-DSCNN) retains the identical memory footprint and computational complexity of the baseline DSCNN while delivering substantially higher accuracy and robustness.

The main contributions of this work are summarized as follows:

- We formulate a complete response-based knowledge distillation pipeline that transfers discriminative knowledge from a high-capacity ResNet teacher to a lightweight DSCNN student for keyword spotting.
- We demonstrate that the distilled student model achieves 96.40% test accuracy on a multilingual crime-related keyword dataset while matching the baseline DSCNN in both memory consumption (≈132 KB) and MAC operations (1,728).
- We provide a comprehensive empirical analysis—covering accuracy, precision, recall, F1-score, false-positive and false-negative rates, and confusion matrices—that quantifies the generalization benefits of knowledge distillation under severe data constraints.
- We show that the proposed approach offers an attractive accuracy–efficiency trade-off, making it particularly suitable for always-on, low-power keyword spotting applications.

II. RELATED WORK

A. Small-Footprint Keyword Spotting

The demand for always-on voice interfaces on resource-constrained devices has driven extensive research into small-footprint keyword spotting (KWS). Early neural approaches relied on fully connected deep neural networks (Chen et al., 2014), which were later surpassed by convolutional architectures that better exploit the time-frequency structure of speech features.

Among convolutional models, the Depth wise Separable Convolutional Neural Network (DSCNN) has proven particularly effective. By decomposing a standard convolution into a depth wise convolution followed by a point wise convolution, DSCNN substantially reduces both the number of parameters and the computational cost while preserving competitive accuracy. Zhang et al. (2017) systematically compared multiple architectures for microcontroller deployment and identified DSCNN as offering the best accuracy–efficiency trade-off. Subsequent studies confirmed these findings: Sørensen et al. (2020) successfully implemented a quantized DSCNN on a Cortex-M4 microcontroller, while Menglong and Xiao-Lei (2020) further improved performance by integrating squeeze-and-excitation blocks into a depth wise separable residual network.

Other efficient designs, such as Temporal Convolutional Networks and various residual or attention-based compact models, have also been explored. Nevertheless, pure DSCNN architectures remain among the most widely adopted solutions for edge KWS because of their simplicity and low memory footprint.

B. Knowledge Distillation in Speech and Keyword Spotting

Knowledge distillation (KD) has emerged as a powerful technique for transferring the generalization capability of large models to compact ones (Hinton et al., 2015). In the response-based formulation, the student is trained to match the softened output probability distribution of the teacher in addition to the ground-truth labels.

In the speech domain, KD has been applied to both automatic speech recognition and keyword spotting. Tan et al. (2021) demonstrated that distillation can significantly improve small-footprint KWS models. Lei et al. (2022) combined DSCNN with a feature-attention mechanism and distillation learning for Chinese wake-word detection, achieving strong accuracy under strict power and latency constraints. Song et al. (2022) further showed that knowledge distilled from large-scale speech recognition models can enhance keyword spotting performance. More recently, Alican et al. (2025) employed a teacher–student framework with self-supervised speech representations to improve few-shot KWS.

These studies confirm that KD is effective for reducing the performance gap between high-capacity and lightweight models. However, most existing work focuses on high-resource languages (primarily English or Chinese) and does not systematically investigate response-based distillation for DSCNN architectures under severe data scarcity.

C. Multilingual and Low-Resource Keyword Spotting

Developing KWS systems for low-resource languages presents additional challenges due to limited annotated data and complex phonetic or tonal characteristics. Multilingual and cross-lingual approaches that leverage pretrained speech representations have recently gained attention (Xie et al., 2023). Self-supervised learning methods have also been explored to alleviate labelled-data requirements.

Despite this progress, research specifically targeting African languages remains scarce. Nigerian languages such as Hausa, Igbo, and Yoruba are tonal and currently lack large-scale annotated keyword datasets. Consequently, the joint application of knowledge-distillation-enhanced DSCNN architectures to multilingual crime-related keyword spotting in these languages has received little attention.

D. Research Gap

While DSCNN provides an efficient backbone for edge deployment and knowledge distillation has proven beneficial for model compression, the combination of response-based KD with DSCNN for low-resource, multilingual KWS—particularly in African tonal languages—remains underexplored. Existing studies rarely address the simultaneous constraints of limited training data, edge-device efficiency, and multilingual capability. This paper fills the gap by developing and rigorously evaluating a response-based knowledge distillation framework that transfers knowledge from a ResNet teacher to a DSCNN student for crime-related keyword spotting in Hausa, Igbo, and Yoruba.

III. PROPOSED METHOD

A. Overall Framework

The proposed system follows a classical teacher–student knowledge distillation paradigm. A high-capacity Residual Network (ResNet) is first trained to convergence on the labelled keyword dataset and then frozen. A compact Depthwise Separable Convolutional Neural Network (DSCNN) is subsequently trained under the joint supervision of the ground-truth labels and the softened posterior distributions produced by the teacher. At inference time, only the student model is retained.

B. MFCC Feature Extraction

Raw audio waveforms are resampled to 16 kHz mono. Pre-emphasis, framing, and windowing are applied, followed by the Short-Time Fourier Transform. Mel-filterbank energies are computed and transformed via the Discrete Cosine Transform to obtain 40-dimensional Mel-Frequency Cepstral Coefficient (MFCC) vectors. Feature matrices are padded or truncated to a fixed temporal length before being fed to the networks.

C. ResNet Teacher Architecture

The teacher is a residual convolutional network that employs skip connections to facilitate the training of deeper layers. Residual blocks of the form $y = F(x) + x$ are stacked to learn

rich hierarchical representations of the MFCC input. The final layer produces logits that are converted into soft targets using a temperature-scaled softmax.

D. Distilled DSCNN Student Architecture

The student network replaces standard convolutions with depthwise separable convolutions. Each depthwise separable block consists of: (i) a depthwise convolution that applies a single spatial filter per input channel, (ii) a pointwise (1×1) convolution that combines the channels, and (iii) batch normalisation and ReLU activation after each convolution. Global average pooling is applied after the final convolutional stage, followed by a fully-connected layer with softmax activation. The resulting architecture contains only approximately 3,741 parameters and requires 1,728 MAC operations per inference.

E. Response-Based Knowledge Distillation Loss

The student is optimised with a weighted combination of two loss terms:

$$l_{total} = \alpha l_{CE}\left(y, \sigma(z_S)\right) + (1 - \alpha)T^2 \, l_{KL}\left(\sigma\left(\frac{z_T}{T}\right), \sigma\left(\frac{z_S}{T}\right)\right)$$

Where z_S, z_T are the student and teacher logits, σ denotes the softmax function, T is the temperature (set to 5), and α = 0.1. The first term enforces fidelity to the hard labels; the second term transfers the inter-class similarity structure learned by the teacher.

F. Training Procedure and Hyper parameters

Both models are trained with the Adam optimiser (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$). Batch size is fixed at 32. Early stopping based on validation loss is employed. The teacher is trained first to convergence; its weights are then frozen and used to generate soft targets for student training.

IV. EXPERIMENTAL SETUP

A. Dataset Description

A custom multilingual dataset of crime-related keywords in Hausa, Igbo and Yoruba was collected from native speakers, yielding 1,453 original utterances (493 Hausa, 422 Igbo, and 538 Yoruba). Ten keywords were selected per language on the basis of phonetic diversity and sociolinguistic relevance.

B. Data Augmentation Strategy

Each original recording was expanded by a factor of ten through pitch shifting and additive Gaussian noise, producing an effective training set of 2,500 samples. The data were split into training (1,600), development (400) and test (500) subsets using a fixed 64/16/20 ratio.

C. Evaluation Metrics

Performance is reported in terms of overall accuracy, precision, recall, F1-score, false-positive rate (FPR), false-negative rate (FNR), memory footprint (KB) and multiply–accumulate (MAC) operations.

D. Implementation Details

All models were implemented in Tensor Flow. Identical data splits, random seeds and hyper-parameters were used for the baseline DSCNN, the ResNet teacher and the distilled student to ensure a fair comparison.

V. RESULTS AND ANALYSIS

A. Accuracy, Precision, Recall and F1-Score

Table 1 summarises the classification performance of the three evaluated models.

Table 1. Classification performance comparison

Model	Test Accuracy	Precision	Recall	F1-score
DSCNN	55.20%	0.552	0.552	0.552
ResNet	82.40%	0.824	0.824	0.824
D-DSCNN	96.40%	0.964	0.964	0.964

The distilled student outperforms both the baseline and the teacher on every classification metric.

B. Memory Footprint and MAC Operations

Table 2 compares the resource requirements of the three models.

Table 2. Memory footprint and computational cost

Model	Memory	MAC Operations
DSCNN	131.80 KB	1,728
ResNet	1,955.79 KB	2,016
D-DSCNN	131.79 KB	1,728

Knowledge distillation improves accuracy by 41.2 percentage points over the baseline without any increase in model size or computational cost.

C. Confusion Matrix Analysis

The D-DSCNN confusion matrix exhibits near-ideal diagonal dominance with only isolated off-diagonal errors. In contrast, the baseline DSCNN shows widespread confusion among acoustically similar classes, while the ResNet still retains residual errors on short-duration keywords. These observations confirm that response-based distillation successfully transfers refined decision boundaries from the teacher to the compact student.

D. FPR / FNR Comparison

Table 3 reports the false-positive and false-negative rates.

Table 3. False-positive and false-negative rates

Model	FPR	FNR
DSCNN	4.98%	44.80%
ResNet	1.96%	17.60%
D-DSCNN	0.40%	3.60%

The distilled model reduces the false-negative rate by more than 90% relative to the baseline while maintaining an extremely low false-positive rate.

E. Discussion of Generalisation Improvement

The large gap between training and test accuracy observed for the baseline DSCNN indicates severe over fitting. The ResNet teacher learns richer representations and generalises better, yet remains too large for edge deployment. Response-based distillation successfully transfers the teacher's decision boundaries to the compact student, yielding a model that simultaneously achieves the highest accuracy and the lowest resource consumption. This result validates the central hypothesis of the present work: knowledge distillation can effectively close the performance gap between lightweight and high-capacity models under severe data constraints without any increase in inference cost.

F. Discussion of Generalisation Improvement

The large gap between training and test accuracy observed for the baseline DSCNN indicates severe over fitting. The ResNet teacher learns richer representations and generalises better, yet remains too large for edge deployment. Response-based distillation successfully transfers the teacher's decision boundaries to the compact student, yielding a model that simultaneously achieves the highest accuracy and the lowest resource consumption. This result validates the central hypothesis of the present work: knowledge distillation can effectively close the performance gap between lightweight and high-capacity models under severe data constraints without any increase in inference cost.

VI. CONCLUSION

This paper presented a response-based knowledge distillation framework that transfers discriminative knowledge from a high-capacity ResNet teacher to a lightweight Depth wise Separable CNN student for multilingual keyword spotting. On a challenging low-resource dataset of crime-related keywords in Hausa, Igbo and Yoruba, the resulting D-DSCNN achieves 96.40% test accuracy while retaining the identical memory footprint (≈132 KB) and computational cost (1,728 MACs) of the baseline DSCNN.

The results demonstrate that knowledge distillation can effectively close the performance gap between compact and high-capacity models under severe data constraints, delivering an attractive accuracy–efficiency trade-off suitable for always-on edge devices. Future work will investigate robustness under realistic noise conditions, language-specific fine-tuning, and the incorporation of feature-based distillation signals.

REFERENCES

[1] Alican Gok, Oguzhan Buyuksolak, Osman Erman Okman, and Murat Saraclar. (2025). Enhancing few-shot keyword spotting performance through pre-trained self-supervised speech models. arXiv preprint arXiv:2506.17686.

[2] Bovbjerg, H. S. (2022). Improving label-deficient keyword spotting through self-supervised pretraining. In Proceedings of the ICASSP 2023 SASB Workshop.

[3] Bushur, J., and Chen, C. (2023). Neural network exploration for keyword spotting on edge devices. Future Internet, 15(6), 1–21.

[4] Chen, G., Parada, C., and Heigold, G. (2014). Small-footprint keyword spotting using deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4087–4091.

[5] Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. arXiv preprint arXiv:1610.02357.

[6] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[7] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[8] Inuwa-Dutse, I. (2025). NaijaNLP: A survey of Nigerian low-resource languages. arXiv preprint arXiv:2502.19784.

[9] 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.

[10] 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.

[11] Menglong Xu and Xiao-Lei Zhang. (2020). Depthwise separable convolutional ResNet with squeeze-and-excitation blocks for small-footprint keyword spotting. arXiv preprint arXiv:2004.12200.

[12] Song, Y., Luo, J., and Li, Y. (2022). Distilling knowledge from large-scale speech recognition models for keyword spotting. In Proceedings of Interspeech 2022, pp. 446–450.

[13] 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.

[14] 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.

[15] 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.

[16] Zhang, Y., Suda, N., Lai, L., and Chandra, V. (2017). Hello Edge: Keyword spotting on microcontrollers. arXiv preprint arXiv:1711.07128.