Towards Efficient Audio-Text Keyword Spotting: Quantization and Multi-Scale Linear Attention with Foundation Models

Rahothvarman P and **Radhika Mamidi**
IIIT Hyderabad
rahothvarman.p@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

Open Vocabulary Keyword Spotting is essential in numerous applications, from virtual assistants to security systems, as it allows systems to identify specific words or phrases in continuous speech. In this paper, we propose a novel end-to-end method for detecting user-defined open vocabulary keywords by leveraging linguistic patterns for the correlation between audio and text modalities. Our approach utilizes quantized pre-trained foundation models for robust audio embeddings and a unique lightweight Multi-Scale Linear Attention (MSLA) network that aligns speech and text representations for effective cross-modal agreement. We evaluate our method on two distinct datasets, comparing its performance against other baselines. The results highlight the effectiveness of our approach, achieving significant improvements over the Cross-Modality Correspondence Detector (CMCD) method, with a 16.08% increase in AUC and a 17.2% reduction in EER metrics on the Google Speech Commands dataset. These findings demonstrate the potential of our method to advance keyword spotting across various real-world applications.

1 Introduction

Keyword spotting (KWS) (López-Espejo et al., 2021) is essential for voice-driven interactions on edge devices, particularly with the growing demand for personalized voice assistants. Traditional KWS (Sainath and Parada, 2015) relies on pre-defined keywords, whereas User Defined Keyword Spotting (UDKWS) (Gurugubelli et al., 2024) enables recognition of user-specific keywords not seen during training, providing more flexibility.

Many UDKWS approaches utilize Query By Example (QbyE) methods (Lugosch et al., 2018; Kim et al., 2019a; Huang et al., 2021), which are limited by variations in speakers and environments. In contrast, text-based keyword enrollment (Sacchi et al., 2019; Shin et al., 2022) methods have gained popularity, as they avoid these issues by focusing on linguistic features which are less prone to variability rather than acoustic characteristics.

End-to-end text enrollment KWS methods (Shin et al., 2022) aim to map audio and text representations to a shared space, but challenges remain in distinguishing similar pronunciations. Self-Supervised Learning speech models (SSLs) (Mohamed et al., 2022) like Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), and Whisper (OpenAI, 2022) offer robust representations but can be computationally expensive for resource-limited devices. Model quantization (Yeh et al., 2022) helps reduce the size and memory requirements of these models, enabling efficient deployment without sacrificing performance.

The attention mechanisms of the transformer architecture (Vaswani et al., 2017) enables models to attend to relevant portions of input sequences. Recent research has explored cross-modal matching techniques, such as attention-based models (Shin et al., 2022), for aligning representations from different modalities such as speech and text effectively. EfficientViT (Cai et al., 2023), introduces a multi-scale linear attention (MSLA) mechanism for vision transformers to reduce the computational complexity of traditional softmax based self-attention. By attending to multiple scales and approximating attention scores, MSLA enables aggregation of fine-grained and coarse-grained features through linear scaling while maintaining low memory usage, making it ideal for deployment on edge devices. In a nutshell, our contributions include:

- Proposing a strategy that enhances foundation model embeddings for speech and text via trainable components atop these models.
- Comparing various foundation models to determine the most effective for this task.
- Applying model quantization to reduce memory and latency for resource-constrained envi-

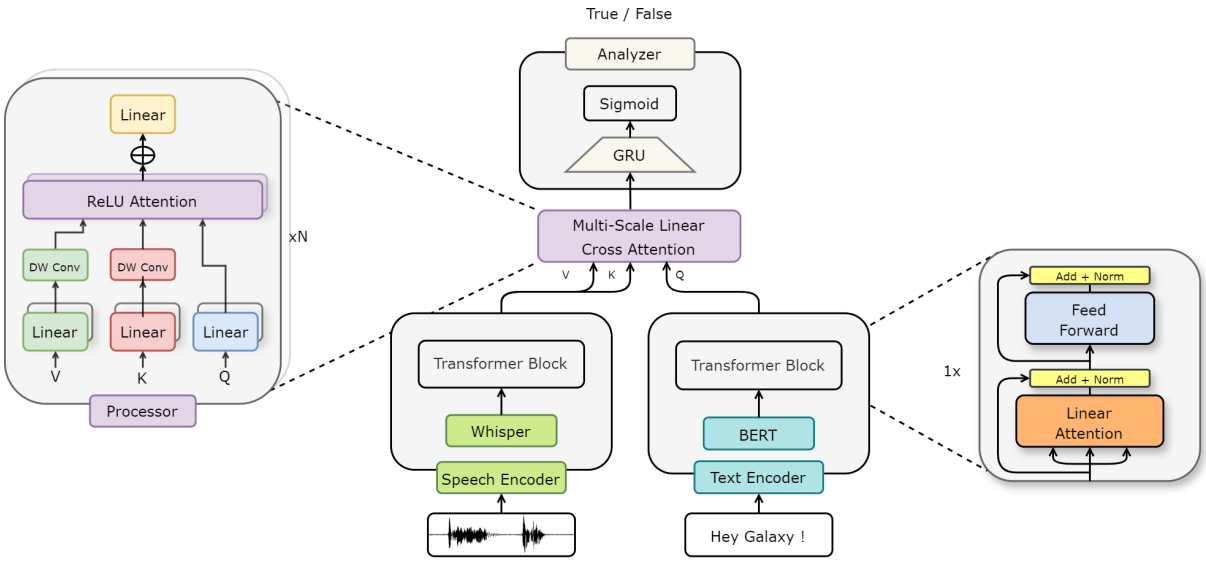

Figure 1: Architecture of the proposed model (MSLA)

ronments.

- Adapting the MSLA mechanism to capture speech-text similarities.
- Demonstrating the robustness of our approach across both closed and open vocabulary scenarios.

2 Proposed Method

This section presents our proposed method, a multi-scale linear cross-attention that computes the similarity between the speech and text by utilizing the capabilities of quantized foundation models. The fundamental blocks of this method are the encoders, a processor and an analyzer. Figure 1, illustrates the general architecture.

2.1 Cross-Modal Speech/Text Encoders

We employed quantized versions of the foundation models such as Wav2Vec2.0, HuBERT, WavLM, and Whisper models as speech encoders and BERT (Devlin et al., 2019) as a text encoder. The motivation of using foundation models for open-vocabulary keyword detection is because of their pre-trained abilities, which allow them to generalize and handle unknown and unseen keywords better. In order to further aggregate the features from these models, we also route the embedding via a transformer block to embed these speech and text features into a shared latent space. The speech and text embedding are denoted as $E_s \in R^{N_s x m}$ and $E_t \in R^{N_t x m}$, respectively, where m represents the embedding dimension and N_s specifies the number of frames and N_t, the tokens in the text, respectively. In all our experiments, we had used the base versions of the foundation models with an output dimension of 768 except for Whisper Base which had 512 as its output dimension.

2.2 Processor

One effective technique for managing cross-modalities has been shown to be cross attention (Vaswani et al., 2017). Hence, the multi-headed cross attention is the foundation of the pattern processor. The network is given the speech embedding E_s as Key K and value V, and the text embedding E_t as query Q. The vanilla cross attention is defined as follows :

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{m} e^{x_j}},\ \sigma(x_i) \in (0,1)^m \quad (1)$$

$$Sim(Q,K) = \sigma(\frac{QK^T}{\sqrt{d_k}}),\ d_k = \frac{m}{n_{heads}} \quad (2)$$

$$Attn = Sim(Q,K) * V \quad (3)$$

The Multi-Scale Linear Attention uses depth-wise separable convolutions with GELU activations for Key K and Value V, avoiding hardware-intensive softmax. While ReLU lacks a non-linear similarity function, it can be addressed by routing Key K and Value V projections through tiny depth-wise separable kernel convolutions. A two-branch design, with 3x3 for K and 5x5 for V token aggregation, captures local dependencies, while the ReLU activated multi-headed attention network handles global dependencies. The proposed MSLA

network is defined as follows :

$$ReLU(x) = max(x, 0) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

$$Sim(Q, K) = ReLU(Q) * ReLU(K)^T \quad (5)$$

$$LinearAttn = Norm(Sim(Q, K)) * V \quad (6)$$

This block creates a similarity map $\in R^{N_t \text{x} N_s}$ between audio and text and builds an attention matrix $\in R^{N_t \text{x} m}$ over this, to be processed by the analyzer.

2.3 Analyzer

In order to examine whether the speech and text are in accord, this block analyzes the attention matrix. It features a Bi-Directional Gated Recurrent Unit (Bi-GRU) and atop, a sigmoid layer for binary class prediction. The final time-steps of the forward and backward GRUs are concatenated and fed to the sigmoid ($\sigma(x) \in (0, 1)$) layer to make predictions.

3 Experimental setup

3.1 Datasets

We conducted experiments using the publicly available Google Speech Commands and Qualcomm Keyword Speech benchmark datasets. The Google Speech Commands v2 (**G**) dataset (Warden, 2018) includes around 100k single-word utterances from 30 speakers, with a training split of approximately 85k samples and a validation split of 10k samples. In addition to the 5k audio snippets from the Google test set, we utilized the Qualcomm Keyword Speech (**Q**) dataset (Kim et al., 2019b), which contains about 4k samples of four commands—'Hey Android', 'Hey Snapdragon', 'Hi Galaxy', and 'Hi Lumina'—recorded under various conditions with 50 speakers and background noise, to assess the model's generalizability beyond standard keywords.

3.2 Training

Open-vocabulary KWS necessitates dealing with a vast number of potential keywords unseen during training. Therefore, for each positive sample containing an audio clip with its corresponding keyword, we required to create a set of negative samples. These negative samples represent audio clips that do not contain the target keyword. The quality and quantity of negative samples significantly impact model performance, as shown in Table 1 of (Shin et al., 2022).

Negative examples were required to enhance the capabilities of our model, as the dataset only included positive samples—that is, audio clips in which the target phrase was present. Five negative samples were randomly generated for each positive sample in the dataset by selecting audio clips that did not contain the target term. We examined the performance of the 16-bit quantized versions of off-the-shelf Speech SSLs as speech encoders. For our text encoder, we chose BERT base.

3.3 Implementation Details

The models were trained on a single NVIDIA RTX 2070 Super Max-Q GPU for 100 epochs, using a batch size of 64. We employed the Adam optimizer with a learning rate of 0.001 and Binary Cross Entropy (BCE) loss. Early stopping was used to prevent overfitting, and the best model was selected based on its performance on the validation set.

4 Results and Discussion

We trained our proposed approach on the Google Speech Commands dataset and evaluated it on the test sets of Google Speech Commands (**G**) and the Qualcomm Keyword Speech (**Q**), which included Out-Of-Vocabulary keywords. The evaluation metrics used were Equal Error Rate (**EER**) and Area Under the Curve (**AUC**) (Table 1). Additionally, we also conducted an ablation study to assess the impact of various architectural and design choices.

Method	EER (%)		AUC (%)	
	G	Q	G	Q
CTC (Lugosch et al., 2018)	31.65	18.23	66.36	89.69
Attention (Huang et al., 2021)	14.75	49.13	92.09	80.13
Triplet (Sacchi et al., 2019)	35.60	38.72	71.48	66.44
CMCD (Shin et al., 2022)	27.25	12.15	81.06	94.51
MSLA	**10.05**	**11.37**	**97.14**	**95.61**

Table 1: Experimental results of our proposed approach MSLA against various baselines

Our approach significantly outperformed baseline methods, even when the training and test datasets differed. This highlights not only the strong generalization capabilities of the underlying foundation models but also the effectiveness of the MSLA mechanism, making it well-suited for open-vocabulary keyword spotting tasks.

Need for Multi-Scale Linear Attention : In our proposed approach, the hardware-friendly Multi-Scale Linear Attention replaces the conventional

Scaled Dot-Product Attention. This mechanism employs ReLU-based attention and utilizes depth-wise separable convolution layers for the Key and Value matrices. By leveraging these convolutions, it generates denser attention maps compared to the vanilla attention network, despite using a linear ReLU function, as illustrated in Figure 2. This design not only accelerates processing but also maintains competitive performance, making it a faster and a more resource-efficient alternative to the conventional attention mechanisms.

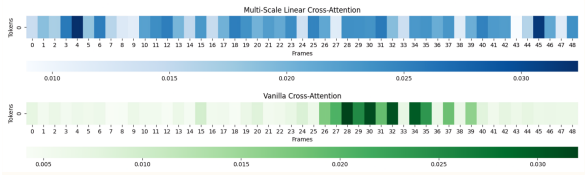

Figure 2: Comparison of Attention Maps from Multi-Scale Linear Attention and Vanilla Attention mechanisms for the positive sample "down"

Role of Model Quantization : Quantizing speech models facilitates efficient deployment in resource-constrained environments with minimal performance loss. Early models were large (300–380 MB) and operated at 32-bit precision; however, quantization can reduce their size by up to 75%, enhancing inference latency, portability, and scalability. To strike a balance between accuracy and latency, we used 16-bit versions. Variations in model sizes and latencies are summarized in Table 2.

Model	32-bit		16-bit		8-bit	
	Size	Lat.	Size	Lat.	Size	Lat.
HuBERT Base	360	100	180	62	90	36
Wav2Vec2.0 Base	360	110	180	65	90	40
WavLM Base	360	130	180	80	90	50
Whisper Base	280	170	140	110	70	65

Table 2: Model Sizes (in MB) and Latencies (in ms) across different Quantization levels in encoding an audio clip of 2s duration

Addition of Fully-Trainable Components : Foundation model embeddings are task-agnostic. To enhance these embeddings for our specific task and tailor them to our needs, we added a GRU layer on top of the foundation models. This enabled knowledge distillation, as the GRU layer was trained concurrently with the primary objective. We focused on the embedding from the final time-step of the GRU layer. After dimensionality reduction, we visualized these vectors and observed dense and distinct clustering based on linguistic and acoustic similarities, which was superior to mean-pooled residual Speech SSL embeddings, seen in Figure 3.

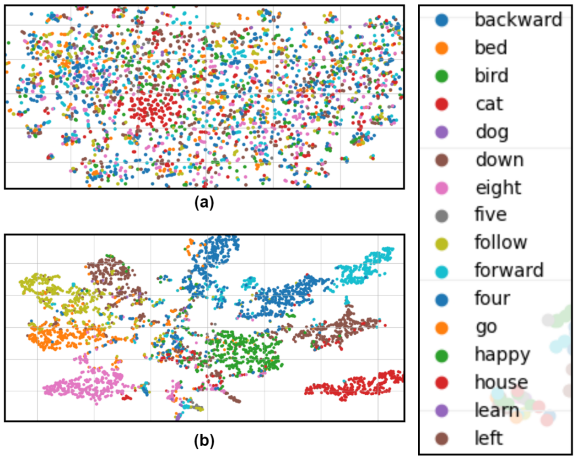

Figure 3: t-SNE Projections; a) Mean Pooled Residual Embeddings (b) Last time-steps of GRU representations

We subsequently replaced GRU with a single transformer block after observing improved performance in terms of the metrics and latency.

Comparison of Foundation Models : We compare several speech foundation models, including HuBERT, Wav2Vec2.0, WavLM, and Whisper, on the Google Speech Commands test split (G). We summarise the results in the Table 3. Whisper achieves the best results, particularly in Recall (99.04%), F1 (93.79), and AUC (97.14%), due to its training on multi-lingual, multi-domain data, allowing it to generalize effectively. HuBERT performs the worst, with lower AUC (89.42%) and F1 (87.36). This can be attributed to its focus on masked prediction rather than downstream tasks like speech-text alignment. WavLM and Wav2Vec2.0 perform competitively, with WavLM slightly ahead due to its multi-task training, leading to better recall and F1. Wav2Vec2.0 excels in precision but falls behind in recall. The performance differences highlight the impact of each model's architecture and training objectives on speech-text alignment tasks.

Encoder	Precision	Recall	F1	AUC
HuBERT	80.23	95.89	87.36	89.42
Wav2Vec2.0	**92.12**	86.61	89.28	89.78
WavLM	88.21	96.69	92.26	93.19
Whisper	89.08	**99.04**	**93.79**	**97.14**

Table 3: Comparison of MSLA under various speech encoders

5 Conclusion

In this study, we proposed the use of quantized self-supervised learning (SSL) foundation models and multi-scale linear attention (MSLA) for an end-to-end user defined keyword spotting (UD-KWS). Model quantization significantly reduces the memory footprint of the foundation models, making them suitable for deployment in resource-constrained environments. Our lightweight Multi-Scale Linear Attention method effectively combines information from speech and text modalities by capturing both local and global dependencies. We compared our proposed approach to state-of-the-art methods using various benchmark datasets and training procedures. Experimental results demonstrated that our proposed strategy significantly outperformed baseline methods, achieving promising results in open-vocabulary keyword detection measures.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Renjie Cai, Menglin Zhang, Yutong Chen, Xiaohan Ding, Tianlong Xia, Jiawei Fang, Jinyang Huang, Zhenyu Liu, Yang Liu, Li Yuan, et al. 2023. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12874–12884.

Sanyuan Chen, Chengyi Wang, Ziqiang Chen, Yu Wu, Shujie Chen, Jian Liu, Kaizhi Yao, Jinyu Zhang, Ming Zhou, and Yanmin Li. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Krishna Gurugubelli, Sahil Mohamed, and Rajesh Krishna K S. 2024. Comparative study of tokenization algorithms for end-to-end open vocabulary keyword detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12431–12435.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jinmiao Huang, Waseem Gharbieh, Han Suk Shim, and Eugene Kim. 2021. Query-by-example keyword spotting system using multi-head attention and soft-triple loss.

Byeonggeun Kim, Mingu Lee, Jinkyu Lee, Yeonseok Kim, and Kyuwoong Hwang. 2019a. Query-by-example on-device keyword spotting. In *Proc. ASRU*, pages 532–538. IEEE.

Byeonggeun Kim, Mingu Lee, Jinkyu Lee, Yeonseok Kim, and Kyuwoong Hwang. 2019b. Query-by-example on-device keyword spotting.

Loren Lugosch, Samuel Myer, and Vikrant Singh Tomar. 2018. Donut: Ctc-based query-by-example keyword spotting.

Iván López-Espejo, Zheng-Hua Tan, John Hansen, and Jesper Jensen. 2021. Deep spoken keyword spotting: An overview.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

OpenAI. 2022. Whisper: Robust speech recognition via large-scale weak supervision. Available at: `https://openai.com/research/whisper`.

Niccolò Sacchi, Alexandre Nanchen, Martin Jaggi, and Milos Cernak. 2019. Open-vocabulary keyword spotting with audio and text embeddings. In *Interspeech 2019*, pages 3362–3366.

Tara N. Sainath and Carolina Parada. 2015. Convolutional neural networks for small-footprint keyword spotting. In *Interspeech 2015*, pages 1478–1482.

Hyeon-Kyeong Shin, Hyewon Han, Doyeon Kim, Soo-Whan Chung, and Hong-Goo Kang. 2022. Learning Audio-Text Agreement for Open-vocabulary Keyword Spotting. In *Proc. INTERSPEECH*, pages 1871–1875.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition.

Ching-Feng Yeh, Wei-Ning Hsu, Paden Tomasello, and Abdelrahman Mohamed. 2022. Efficient speech representation learning with low-bit quantization.