Whisper-tiny 모델을 활용한 음성 분류 개선: 확장 가능한 키워드 스팟팅 접근법

시바니 산제이 콜레카르, 진현석, 김경백
전남대학교 인공지능학과
shivanikolekar@gmail.com, ggyo003@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

Enhancing Speech Recognition with Whisper-tiny Model: A Scalable Keyword Spotting Approach

Shivani Sanjay Kolekar, Hyeonseok Jin, Kyungbaek Kim
Dept. of Artificial Intelligence Convergence,
Chonnam National University
Gwangju, South Korea

Abstract

The effective implementation of advanced speech recognition (ASR) systems necessitates the deployment of sophisticated keyword spotting models that are both responsive and resource-efficient. The initial local detection of user interactions is crucial as it allows for the selective transmission of audio data to cloud services, thereby reducing operational costs and mitigating privacy risks associated with continuous data streaming. In this paper, we address these needs and propose utilizing the Whisper-Tiny model with fine-tuning process to specifically recognize keywords from google speech dataset which includes 65000 audio clips of keyword commands. By adapting the model's encoder and appending a lightweight classification head, we ensure that it operates within the limited resource constraints of local devices. The proposed model achieves the notable test accuracy of 92.94%. This architecture demonstrates the efficiency as on-device model with stringent resources leading to enhanced accessibility in everyday speech recognition applications.

1. Introduction

The proliferation of voice-enabled interfaces across consumer electronics and industrial applications necessitates the development of Advanced Speech Recognition (ASR) systems that are not only accurate but also computationally efficient and privacy-preserving. These systems (figure 1) are pivotal for facilitating intuitive human-computer interactions through voice commands, thereby enhancing user accessibility and operational efficiency. However, the real-time processing requirements of these systems, coupled with the widespread deployment in resource-constrained environments, present significant challenges in terms of scalability, responsiveness, and data security [1].

Speech recognition, particularly keyword spotting which involves the recognition of predetermined phrases from streaming audio, remains a critical task in ASR, demanding immediate responsiveness and high precision under tight computational budgets, particularly on edge devices such as smartphones and IoT devices. Additionally, the imperative to process data on-device to mitigate latency and protect user privacy complicates the reliance on centralized cloud computing resources, emphasizing the need for localized processing solutions [2]. Traditionally, large neural network models have dominated the field of ASR due to their superior

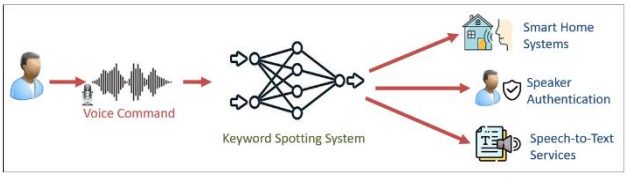

Figure 1: Keyword Spotting System example applications.

ability to model complex linguistic patterns and diverse acoustic conditions. However, these models often require significant computational resources that are not feasible for continuous operation on edge devices [3].

Most keyword recognition models do not fully exploit the advanced capabilities of latest neural network architectures that have been developed for purposes other than speech recognition, which could potentially optimize ASR performance metrics. These advanced architectures, incorporating techniques such as attention mechanisms and transformer models, have demonstrated significant potential in enhancing processing efficiency and model adaptability across varied applications [3].

We propose a methodology for keyword spotting based on Whisper-tiny model with a lightweight classification head. Our proposed method focuses on performing the contextual

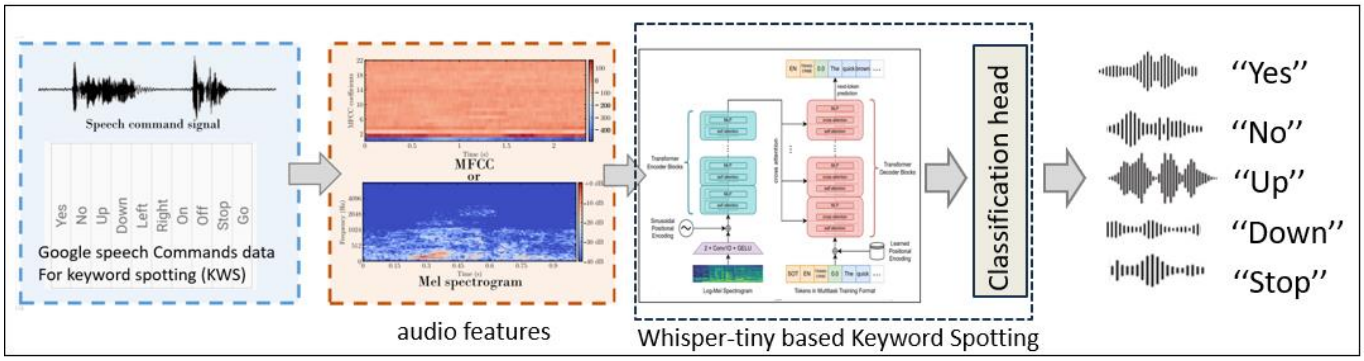

Figure 2: General Architecture for Whisper-tiny based Keyword Spotting (KWS) Model

speech recognition tasks on local devices with limited resources without compromising the performance accuracy.

2. Related Works

The landscape of advanced speech recognition systems has been rapidly evolving, particularly in the realm of keyword spotting (KWS), which offers a specialized, resource-efficient alternative to full-sentence transcription systems. Recent studies have illuminated the advantages of KWS, such as faster response times and reduced computational demands, making it ideal for voice-activated interfaces and real-time command recognition [3],[7]. Moreover, KWS systems focus on specific trigger phrases, thus enhancing power efficiency and privacy. Existing on-device KWS approaches have either simplified neural architectures or compressed larger models, with the former often losing accuracy in noisy settings, and the latter not always meeting the tight constraints of edge devices [2]. Advanced neural network architectures using attention mechanisms and transformers offer promising enhancements in efficiency and adaptability across various applications [4].

In this paper, we propose a keyword spotting (KWS) method based on Whisper-tiny model as shown in figure 2. Utilizing Whisper-tiny along with classification head for ASR systems directly addresses previous challenges by leveraging a more compact model that is inherently suited for on-device applications, thus mitigating the heavy resource demands of larger models.

Its optimized framework ensures efficient processing without compromising on performance, even in noisy environments, overcoming the accuracy issues faced by overly simplified models. Additionally, Whisper-tiny incorporates advanced neural network techniques, such as attention mechanisms, that enhance its adaptability and robustness across diverse acoustic conditions. These features make it an ideal choice for edge devices, balancing the need for reducing energy consumption, efficient device communication and maintained privacy with advanced speech recognition capabilities.

3. Keyword Spotting (KWS) Model based on Whisper-tiny

Whisper-tiny based KWS Architecture: Whisper is a Transformer based encoder-decoder model, also referred to as a sequence-to-sequence model. It was trained on 680k hours of labelled speech data annotated using large-scale weak supervision [6].

The Whisper Processor is used to pre-process the audio inputs by converting them to log-Mel spectrograms for the model and to post-process the model outputs by converting them from tokens to text.

The self-attention mechanisms within the whisper transformer architecture enable the model to dynamically focus on different parts of the audio input sequence during training and inference, thereby focusing on influential data of keywords being spotted. The embedding layer maps the input features to a higher-dimensional vector space. This embedding space is optimized during training to facilitate the model's discrimination between different keywords.

Whisper-tiny model represents a compact version of the Whisper architecture designed for efficient inference with minimal computational resources, while maintaining competitive performance [6].

We instituted a consistent random seed to ensure the stability of our results across multiple runs. Furthermore, we specified the precision level for matrix multiplication operations, reinforcing the exactitude of our computational process. During our training loop, detailed logging was done, providing the model's performance and behavior during each epoch.

Finally, to provide accessibility and further inquiry, we established a systematic protocol for model state preservation. We saved the model checkpoints with high precision and performance metrics annotations.

Model performance was checked using accuracy and loss metrics, with the former serving as the primary indicator for saving model checkpoints. Post-training, the model with the highest validation accuracy was evaluated on the test set to provide an unbiased assessment of its generalization capabilities.

Data augmentation and sampling strategy: Acknowledging the potential for class imbalance, which is a common challenge in speech recognition tasks, we augmented the training set with synthesized silent background audio samples. This strategy aimed to balance the distribution of classes within the dataset, thus circumventing the model's predisposition towards more frequently occurring classes.

The sampling technique includes WeightedRandomSampler,

which guarantees an equitable representation of each class during the training phase. This approach negates any inherent bias by enabling equal learning opportunities for each class, ensuring no single class disproportionately influences the model's learning trajectory.

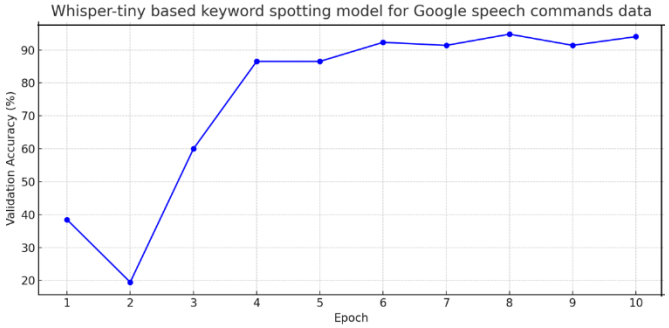

Figure 3: Whisper-tiny based keyword spotting model for Google speech commands data.

4. Evaluation

Dataset Description: The input for our keyword spotting system comprises audio recordings of spoken commands. These commands were sourced from the Google Speech Commands dataset, which is a widely acknowledged benchmark collection for training keyword spotting models. The dataset encompasses a diverse set of spoken commands recorded under various acoustic conditions, ensuring a robust training process for our model [5].
We use the latest dataset version (v0.02) containing more than 30 different single line command keywords. From these, 10 keywords are used as commands by convention: "Yes", "No", "Up", "Down", "Left", "Right", "On", "Off", "Stop", "Go" and other words are considered to be auxiliary. Their function is to teach a model to distinguish core words form unrecognized ones.
It also includes a set of longer audio clips that are either recordings or a mathematical simulation of noise.

The keyword spotting model based on pretrained Whisper-tiny model was evaluated for accuracy and loss metrics for 10 epochs and 0.001 learning rate.
The validation performance was observed as 94.04 % and loss of 0.0047. The performance after each epoch was consistently increased, which can be observed in figure 3. The test accuracy was observed as 92.94 % with a loss value of 0.0064. The experiment was conducted on a server while allowing only limited resources to use while performing the training updates with incredibly lightweight classification head of 780 thousand parameters. Total time taken for training completion was approximately 40 minutes for 10 epochs.

5. Conclusion and Future Work

In this paper, we proposed a keyword spotting method based on the Whisper-tiny model with a lightweight classification head. We evaluated the model with Google speech commands dataset over a limited number of epochs. The proposed method achieved 92.94% test accuracy in just 10 epochs on a local source constrained device in only 40 minutes. The Whisper-tiny based keyword spotting model's impressive adaptability to a small-scale architecture without significant compromise to accuracy demonstrates effective capability for its deployment in edge devices.
In the future, we plan to evaluate the model’s adaptability to distributed data on multiple local training nodes with federated learning. This approach will provide stronger insights into model’s robustness and generalization capability for ASR systems.

Acknowledgement

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-00156287, 50%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629, 50%) grant funded by the Korea government(MSIT).

References

[1] Sharif, Khairunisa, and Bastian Tenbergen. "Smart home voice assistants: a literature survey of user privacy and security vulnerabilities." Complex Systems Informatics and Modeling Quarterly 24 (2020): 15-30.
[2] Cheng, Shitong, et al. "Task offloading for automatic speech recognition in edge-cloud computing based mobile networks." 2020 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2020.
[3] Wang, Siyin, et al. "Can Whisper Perform Speech-Based In-Context Learning?." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.
[4] Lyu, Ke-Ming, Ren-yuan Lyu, and Hsien-Tsung Chang. "Real-time multilingual speech recognition and speaker diarization system based on Whisper segmentation." *PeerJ Computer Science* 10 (2024): e1973.
[5] Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." arXiv preprint arXiv:1804.03209 (2018).
[6] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International Conference on Machine Learning. PMLR, 2023.
[7] Mwase, Christine, et al. "Communication-efficient distributed AI strategies for the IoT edge." Future Generation Computer Systems 131 (2022): 292-308.