**Small-footprint Audio Event Identification with Teacher-Student Deep Learning**

**By**

Tianqi Jiao

Zhe Sun

Xifan Tang

Yuxin Wang

### Supervisor: Anil Chaturvedi, Donald Patchell

A Capstone Project

Submitted to The University of Chicago in partial fulfillment of the requirements for the degree of

Master of Science in Applied Data Science

Division of Physical Sciences

October 2023

**Abstract**

To enhance keyword spotting functionality and the generalization ability on customized keywords for Analog Devices' MAX78000 Chip, a knowledge distillation (Teacher-Student) deep learning framework was implemented with dimensionally reduced pre-trained Wav2Vec2 serving as the teaching model to extract the features of the raw audio. In order to address the integration of an advanced audio recognition system within the limited storage space of the chip, a student model which takes raw audios and learns to mimic the teacher using a simpler Convolutional Neural Network was built. The adaptation yielded a few-shot classification model with an accuracy of 90.23% among keywords detection without background, and 92.57% with background that consist of daily conversations. Model was validated with self-recorded, non-standardized audio data that were generated by a text to speech system, Piper, for the generalizability of this model in customizable keyword identification. An accuracy of 96.96% was achieved with 50 training samples and training time of less than 2 minutes. The last dimensionality reduction layer is the only part that needs to be trained facing new customized keywords, which reduces the training time and the requirement of training sample size significantly. Therefore, the teacher-student model is proven to be accurate in keyword detection and is easily customizable and generalizable for new keywords.

*Keywords:* audio event identification, knowledge distillation, teacher-student deep learning, generic keyword spotting, keyword detection, Wav2Vec2, convolutional neural network, dimensionality reduction, Piper, distance metric learning

**Executive Summary**

Analog Devices addresses the challenge of enhancing audio event detection and keyword spotting within hardware limitations and custom keyword adaptability using its MAX78000 Chip. A Teacher-Student Deep Learning framework has been developed to efficiently train a dimensionality-reduced keyword spotting model suitable for this chip.

The teacher model combines the pre-trained Wav2Vec2 for feature extraction from raw audio, producing a 1024-dimension embedding, with a dimensionality reduction DNN layer that condenses this to 128 dimensions. This model achieved a 92.57% accuracy on a keyword dataset with background noise.The student model, KWSNetv3, integrates audio analysis and dimensionality reduction in a single stage using Maxim’s ‘ai8x’ library, designed for the MAX78000 Chip, and reaches a 128-dimension embedding. Though its accuracy is lower (54.86%) at first due to a simpler CNN, it effectively mimics the teacher model.

A key innovation is the addition of a few-shot dimensionality reduction DNN layer to the student model. This layer requires only 50 training samples per class, significantly reducing training time and adapting efficiently to new, unseen samples. With this enhancement, the model achieved accuracies of over 90% in keyword detection both with and without background noise, proving its effectiveness in complex audio environments with minimal hardware. Validation of generalizability with self-generated data yields an accuracy of 96.96%. This proves that the model is easy generalizable and suited for customizable keyword spotting tasks.

This project not only demonstrates the feasibility of advanced customizable audio event detection in resource-limited systems, but also paves the way for future innovations in embedded auditory technology.

**Table of Contents**

Introduction ................................................................................................................................... 5

Business Problem/Opportunity Statement ......................................................................... 5

Analysis/Research Goals ................................................................................................... 6

Scope …………………. ................................................................................................... 7

Background ……........................................................................................................................... 9

Business Partner Background ........................................................................................... 9

LiteratureReview................................................................................................................ 9

Data............................................................................................................................................... 12

Data Sources .................................................................................................................... 12

Data Description Analysis …........................................................................................... 13

Methodology ................................................................................................................................ 15

Feature Engineering ......................................................................................................... 15

Modeling Frameworks ..................................................................................................... 15

Findings ....................................................................................................................................... 20

Discussion .................................................................................................................................... 34

Conclusion ................................................................................................................................... 37

Reference ..................................................................................................................................... 39

**Introduction**

**Opportunity Statement**

In today's technological landscape, audio technologies have gained significant traction across various industries. Voice-controlled virtual assistants and home automation systems have revolutionized the way people interact with devices, enabling seamless task execution through simple voice commands. To further enhance the capabilities of these technologies, it is imperative to focus on audio event detection and keyword spotting. The field of deep learning-based audio event identification, which encompasses tasks such as keyword detection, speaker identification, and the detection of rare events like glass breaking or gunshots, has witnessed significant growth over the past decade. Despite its widespread application in various industries and everyday scenarios, customization remains a challenge due to the prolonged and energy-intensive training processes.

The challenge lies in developing audio event identification systems that are not only compatible with small embedded chips but also powerful and flexible with a small number of new customized keywords. Current deep learning models are able to accomplish the tasks of identifying keywords or wake words. However, these models are not generalizable and customizable for various target keywords. Furthermore, balancing model size reduction with performance enhancement is also a critical concern in this domain. In light of this, our collaborative effort with Analog Devices, Inc. aims to address these challenges by leveraging the potential of teacher-student deep learning.

The collaboration with Analog Devices, Inc. was established to address these challenges by employing teacher-student deep learning models. ADI initiated a method using a Teacher-Student Model. In this approach, a pre-trained comprehensive model, termed the “teacher”, imparts training to the streamlined Convolutional Neural Network(CNN) models, referred to as the “students''. The objective was to have a model that could discern custom keywords amidst background noise with a minimum accuracy of 90%. This Teacher-Student Model was designed to offer a dimensionally reduced solution for keyword detection challenges, focusing on extracting device-activation keywords from audio segments, fitting the parameter constraints of ADI’s CNN-specific chips. It also provided a pathway for easy customization based on a small amount of user-defined keyword training samples.

Addressing this challenge effectively led to a reduction in the collection or preparation effort of large amounts of new customized keyword training audio samples. It also reduced the training time on new customized keywords by an estimated 70% because the last Distance Metric Learning (DML) layer is the only part that needs to be trained for the new customized keywords, which can also reduce the computational power demand of GPU. Furthermore, the training process of this model only requires a small number of data samples for each new keyword, thus significantly shortened the training time as well as reducing the requirement of data collection. This not only positioned ADI to broaden its footprint in the custom audio detection market but also attracted a new clientele. Consequently, this translated to potential operational cost savings and set the stage for product diversification. The projected revenue growth for Analog Devices can be calculated if the related data becomes available in the future.

**Research Goals**

The objective was the development of an easy customizable audio event identification model that is compatible with the MAX78000 chip with embeddings of less than 128 dimensions; trainable with a small number of data samples, while maintaining a high accuracy (90%+) in keyword detection. An in-depth review was conducted on the prevailing deep learning models used for audio event identification, especially the Hugging Face Wav2Vec2. This review underscored their strengths, potential limitations, and avenues for optimization, laying the groundwork for informed project decisions.

Experiments with publicly available datasets were essential in assessing the adaptability and efficiency of these models across varying audio event identification scenarios. This investigative phase was crucial in training models that would best fit the project's unique requirements.

The most important technology was dimensionality reduction utilizing the evaluation method of distance metric learning. This ensured that the solution was in line with the computational capabilities of the MAX78000 chip. The isolated dimensionality reduction layer was the only part to be trained on new customized keywords, which is the key for faster training, less training sample demand and less computational power demand.

The construction of a student model, is tailored for the MAX78000 chip from Analog Devices and able to generalize on a small number of new customized keyword samples. This model was envisioned to proficiently process quantized audio, underlining the project's commitment to efficiency and adaptability.

**Scope**

The scope of the work had certain predefined constraints. The research approach was specifically tailored to the Qualcomm and Librispeech datasets. This specific focus may restrict the application of the methodology to other datasets or broader audio sources. Audio data was restricted to samples of 1.5 seconds, and the sampling rate was fixed at 16,000 samples per second. This specificity means that the findings might not be directly transferable to audio samples of varying lengths or sampling rates.

While the methodology is detailed and provides significant insights into audio event identification, the deliverable's application is primarily optimized for keyword spotting. Readers should be cautious when considering its use for other audio processing tasks. Dependence on particular tools, such as the Hugging Face Wav2Vec2 model and the MAX 78000 chip, further narrows the scope. Any updates or changes to these tools could necessitate revisions to the methodology.

**Background**

**Business Partner Background**

Analog Devices, Inc. (ADI) stands as a prominent global semiconductor company that specializes in data conversion, signal processing, and power management technology. With the incorporation of Maxim Integrated, ADI expanded its capabilities, now holding a combined value of $68 billion. The technologies pioneered by ADI influence a diverse range of industries, such as communications, computers, automotive, and consumer electronics applications.

The company's analog, mixed-signal, and digital signal processing (DSP) integrated circuits (ICs) are instrumental in converting real-world phenomena—like light, sound, temperature, motion, and pressure—into electrical signals. This conversion is pivotal for the optimal operation of numerous electronic devices. Notably, ADI solutions are integral to smartphones, aiding voice recognition and noise cancellation features. Additionally, they play a crucial role in smart home devices by facilitating energy-efficient and intuitive controls. Beyond these applications, ADI's products serve healthcare, industrial automation, aerospace, and defense sectors.

Innovation remains a cornerstone of ADI's operations. By continuously advancing technological boundaries and prioritizing research and development, ADI ensures its positioning at the cutting edge of the semiconductor industry. The company's emphasis on innovation has resulted in an array of advanced products and solutions. The potential of specialized chips, such as the MAX78000, especially when considered alongside teacher-student models, promises to cater to the rising demand for energy-efficient audio event identification solutions in diverse sectors.

**Literature Review**

The domain of audio processing has witnessed remarkable advancements, particularly in the spheres of keyword spotting and the teacher-student deep learning paradigm. This body of research offers key insights into potential strategies and methodologies that directly inform the progression and choices made in related projects.

Deep learning models for keyword spotting (KWS) have shown promise in modern electronic devices. López-Espejo et al. (2021) emphasize the critical role of deep KWS in functions such as voice assistant activation. Moreover, their study suggests that a strategic approach to these models can lead to optimized computational complexity—a crucial consideration for incorporating KWS technologies in chips challenged by computational limitations.

The concept of Knowledge Distillation, commonly understood as the teacher-student training paradigm, has been illuminated in contemporary studies. As elucidated by Abbasi et al. (2019) and further expounded by Hu et al. (2022), this paradigm centers on the knowledge transfer from an advanced model (the 'teacher') to a more streamlined version (the 'student'). The overarching objective is to align the student model's performance with its more complex teacher counterpart.

The application of this teacher-student paradigm has demonstrated tangible benefits, especially in the area of sound event detection (SED). For instance, Ng et al. (2018) found that this methodology can achieve results comparable to traditional methods in speaker recognition systems, but with the advantage of a reduced model size—making it apt for specialized devices like the MAX78000.

Further exploration into noisy student-teacher training in streaming keyword spotting by Park et al. (2021) unveiled the model's potential to improve accuracy, especially in challenging testing conditions. Such findings hint at the broader applications of these models, such as in wake-up word detection tasks where ambient noise is a concern.

The essence of model compression, especially for on-device keyword spotting, has been articulated by Tucker et al. (2018). Their research underscores the value of harnessing knowledge from an ensemble of models, pointing towards the possibility of enhanced performance without imposing undue computational demands.

Sequence-level teacher-student training also holds significance, as advocated by Wang et al. (2018) in the context of real-time language assessment. Such training methodologies are believed to streamline computational processes.

Furthermore, a prevalent theme in contemporary studies, such as that by Jia et al. (2019), revolves around the adoption of multiple student models. These distributed learning systems, nested within the teacher-student architecture, can potentially fast-track the learning curve by orchestrating simultaneous training across diverse platforms.

In synthesis, the existing literature converges on the transformative potential of keyword spotting and the teacher-student paradigm, underscoring their importance in shaping the future of audio processing.

**Data**

**Data Sources**

Two primary datasets, the LibriSpeech and the Qualcomm keyword speech dataset, were utilized for the study.

The LibriSpeech-100h-clean-audio dataset, a component of the larger LibriSpeech Automatic Speech Recognition (ASR) corpus, originates from the LibriVox initiative. This dataset provides 100 hours of clear English speech data files characterized by minimal background interference and consistent recording conditions. The recordings of this dataset served a dual purpose: they were instrumental in training and evaluating ASR models and also ensured enhancement of speech recognition and keyword spotting capabilities. To further refine the model's efficiency, recordings from this dataset acted as a backdrop during training, preventing unintended activations by commonly used daily phrases.

The Qualcomm Keyword Speech dataset contains 4,270 utterances of select English keywords: "Hey Android," "Hey Snapdragon," "Hi Galaxy," and "Hi Lumina," presented by 50 participants. For the study, the raw audio from this dataset was crucial in generating embeddings essential for the dimensionality reduction model.

While each dataset has its specific purpose and focus, there are some gaps and limitations. Both datasets only include audio recordings in English and are mainly spoken by native speakers, thus it will require an additional training process to transfer the model to instances of other languages or English spoken by non-native speakers. Besides, the LibriSpeech dataset is limited to clean speech and audiobook genres, which makes the training output difficult to be generalized to real-world situations with background noises and other confounding factors. Lastly, the Qualcomm Keyword Speech dataset includes limited vocabulary and is not legally allowed to be directly used in developing products. Therefore, the role of these two datasets is strictly limited to preliminary explorations.

In addition to the two primary datasets, a self-generated, unnormalized dataset was used to validate the generalizability and customizability of this model for new, unseen data that are closer to real-word application cases. This validation dataset was generated using Piper, a fast, local neural text to speech system. There are 4 keywords for which audio data were generated: “good night cortana”, “hello chicago”, “hey jarvis”, and “morning siri”. For each keyword class, 200 audio files were generated using Piper. Each class was separated into a training set (50 samples) and a testing set (150 samples).

**Data Descriptive Analysis**

During the exploration phase, a thorough descriptive analysis was conducted on the datasets. The primary objective of this analysis was to garner insights into the distribution and composition of the data, particularly focusing on the dependent variable: the effectiveness of speech recognition and keyword spotting.

The exploratory data analysis illuminated several key aspects. The LibriSpeech-100h-clean-audio dataset showcased a diverse range of English speakers, evident from the varied pitch, modulation, and tone patterns across the data. The frequency distribution of certain phonemes and their respective occurrences reveal a predilection towards certain phonetic sounds, which is indicative of the dataset's English-centric nature.

On inspecting the Qualcomm Keyword Speech dataset, the distribution of the utterances across the four selected keywords was relatively even. This uniformity in distribution suggests a balanced representation for each keyword, vital for unbiased model training.

It was observed that while the majority of recordings in the LibriSpeech dataset displayed minimal background interference, some recordings showcased minor fluctuations. These perturbations, though minimal, could potentially impact model training.

In terms of data distribution, it is essential to understand the nature of the dependent variable. The analysis demonstrated that the effectiveness of speech recognition and keyword spotting followed a normal distribution, implying that most of the recordings were of average clarity and audibility, with fewer instances of exceptionally clear or muddled recordings.

In summation, the descriptive analysis served to paint a comprehensive picture of the datasets' characteristics. This analytical process highlighted the strengths and potential challenges embedded within the data, preparing the ground for subsequent phases of model training and validation.

**Methodology**

**Feature Engineering**

The methodology's foundation for the audio event identification system was laid with exhaustive feature engineering. The opening phase involved the meticulous preparation of the Qualcomm and Librispeech datasets and the generated Piper dataset.

In the data preparation process, raw audio data, spanning a duration of 1.5 seconds, was sampled at an impressive rate of 16,000 samples per second, yielding 192x128 samples. Owing to the analog nature inherent in audio signals, there was a pivotal transition from continuous to discrete representation achieved via relentless sampling. This transition showcased the sound amplitude at diversified time intervals. To meet the intricate requirements of convolutional neural networks and simultaneously boost computational efficiency, this audio data underwent a transformation into a two-dimensional array. The Qualcomm dataset and Piper dataset saw refinement with the inclusion of specific keywords, and the Librispeech clean test underwent modifications to capture background noises.

A phase that merits special attention was the extraction of standout audio features via the Hugging Face Wav2Vec2 model. Crafted with a convolutional neural network structure, this model was adept at processing raw audio, converting it into representations primed for speech recognition systems. The transformative journey from speech signals to standardized feature vectors played a pivotal role in facilitating the training of classifiers to anticipate the correlating textual content. Following the rigorous extraction process spanning both datasets, a 1024-dimensional embedding for each audio excerpt found its place under the 'wav2vec2' key in the 'test\_features'.

**Modeling Frameworks**

A Teacher-Student Deep Learning framework was developed to train a dimensionality reduced keyword spotting model implemented on the MAX78000 Chip.

The teacher model contains two parts, the pre-trained Wav2Vec2 and a zero-shot dimensionality reduction Deep Neural Network(DNN) layer.

Wav2Vec2 extracted the features from the raw audio, and got a 1024-dimension embedding. Wav2Vec2 is a pretrained model for Automatic Speech Recognition (ASR) and was released in September 2020 by Hugging Face, which is a convolutional neural network (CNN) that takes raw audio as input and computes a general representation that can be input to a speech recognition system. The objective is a contrastive loss that requires distinguishing a true future audio sample from negatives. The core idea of Wav2Vec2 is to convert the speech signal into a series of fixed-length feature vectors, and then use these vectors to train a classifier to predict the text content represented by the speech signal. Using a novel contrastive pre-training objective, Wav2Vec2 learns powerful speech representations from more than 50,000 hours of unlabeled speech. Like BERT's masked language modeling, the model learns contextualized speech representations by randomly masking feature vectors before passing them to a transformer network. Wav2Vec2 was used for feature extraction on both Qualcomm dataset and the Librispeech dataset, obtaining an embedding of 1024 for every audio, which was correspondingly saved in the 'wav2vec2' key of ‘test\_features’ as a baseline for subsequent training of the student model. Due to the large size of the datasets and the complexity of the Wav2Vec2 model, it took about 2 weeks to extract the embeddings of all audios.

The zero-shot dimensionality reduction DNN processes the Wav2Vec2 extracted 1024-dimension embedding into 128 dimensions. Distance Metric Learning(DML) is a key component in this stage and is used to learn how to measure the similarity or distance between different categories. In embedding space, the larger the inner product between two vectors, the smaller the angle, which means the more similar they are. The goal of DML is to enable the model to adapt the distance measures learned during training to new, unknown categories. This can be achieved by embedding samples of known categories into a feature space and then learning how to measure the distance between different categories. Once distance measures are learned, the model can identify new categories or perform other zero-shot learning tasks based on these measures. Zero-shot machine learning is a model trained to recognize and understand classes or concepts it has never encountered during the training phase. In traditional machine learning, models are trained on a fixed set of classes, and they struggle to recognize or classify items that belong to classes outside of their training data. However, zero-shot learning enables models to generalize their knowledge to unseen classes, making it a valuable approach for expanding the capabilities of AI systems. In conclusion, zero-shot dimensionality reduction aims to improve the generalization ability of the model so that it can better adapt to new and unknown situations in the future training or application.

The student model, KWSNetv3, mimicked its teacher, fulfilling the audio analysis and dimensionality reduction in one stage, using the Maxim’s built library ‘ai8x’, which fits the parameters restrictions of the MAX 78000 chip. Finally it got a 128-dimension embedding. Different from Machine Learning training in theory and ideal situations, the Machine Learning model for audio keyword recognition in hardware should have smaller dimensions and fewer parameters. The student model has 10 convolutional layers and a fully connected layer, uses dropout to prevent overfitting, utilizes maximum pooling and average pooling to connect different convolutional layers, and uses ReLU as the activation function. The gradients are computed, and the model's weights are updated using the Adam optimizer. A learning rate scheduler monitors the validation loss, and if the loss plateaus, it reduces the learning rate. The model is saved at regular intervals (every 100 epochs) for future use. The ‘tester’ function is called to assess the model's keyword spotting performance. This function evaluates the embeddings' cosine similarity between the model's generated embeddings and the embeddings generated by the student model.

Then a few-shot dimensionality reduction Deep Neural Network(DNN) layer was added after the student model, referred to as DML2, compressing input vectors from 128 dimensions down to 64 dimensions. A sequential construct, utilizing nn.Sequential, served as the foundation for the model's ability to encode data into a lower-dimensional space while preserving its complex characteristics. The model's architecture, comprising linear and nonlinear layers, was instrumental in learning an enriched representation of the input data. The initial layer, despite maintaining the dimensionality of the data, employed the network's weights and activation functions to extract complex features from the input. The second linear layer was pivotal in reducing the data's dimensions by half. Despite the significant reduction, the design and interplay of layers ensured minimal information loss. In conclusion, the DML2 model successfully executed dimensionality reduction through a sophisticated orchestration of linear and nonlinear processes. This careful configuration did not simply compress data; it restructured the data to highlight and preserve underlying patterns, facilitating accurate similarity evaluations. In the training process, this model explored only 50 samples per class, which extracted useful information from limited samples to make accurate predictions on new, unseen samples. This last dimensionality reduction layer is the only part that needs to be trained facing new customized keywords, which reduces the training time by directly using the model structure and parameters trained by the large training dataset, reduces the demand of a large number of training samples of the new customized keywords and reduces the need of computational power significantly by avoid the complex training on GPU.

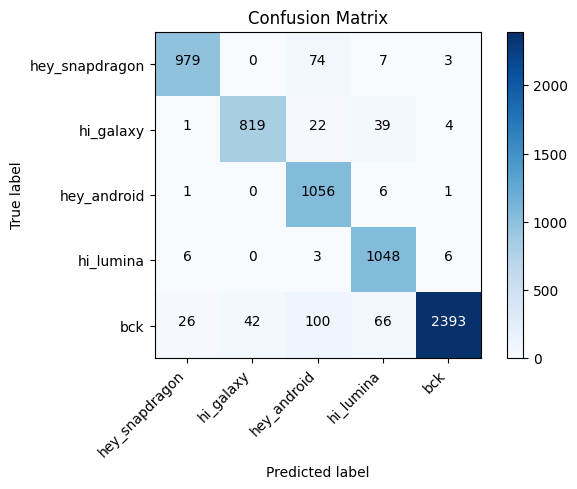
In the methodology for model evaluation and validation, the inner product was utilized rather than cosine similarity or Euclidean distance for distance metric learning. This decision stems from the fact that Analog is a hardware company, and the inner product is an inexpensive operation when executed on a chip. Inner product calculates the sum of the products of corresponding entries of the two sequences of numbers, which is computationally less expensive, especially when used with sparse vectors. By optimizing the distance metric learning with the inner product, the efficiency and performance of the audio event identification system were enhanced on the MAX78000 chip with less training time and training cost.

**Findings**

The training process involves two primary components: the teacher model and the student model. The teacher model utilized the 1024-dimensional embeddings derived from 1.5-second audio samples through Wav2Vec2. Subsequently, it employed a zero-shot dimensionality reduction layer (DRL) to decrease the dimensionality of the embeddings to 128 dimensions. The student model replicated this procedure, producing 128-dimensional embeddings directly from the raw audio data. A few-shot DRL using only 50 samples was then trained to be applied to both the teacher and student models, resulting in further reduced 64-dimensional embeddings and the final classification outcome of the model.

The accuracy of the teacher model, which is Wav2Vec2 followed by a zero-shot DRL, on the background included training dataset was 93.93% and this is also the baseline of the student model. Even without the few-shot training, Wav2Vec2 can also get a very high accuracy on background included dataset, which is because it is a well-developed and well-trained mature self-supervised model. The corresponding confusion matrix was plotted to visualize the classification result.

**Figure 1.** *Confusion matrix for keywords and background before DML in teacher model*

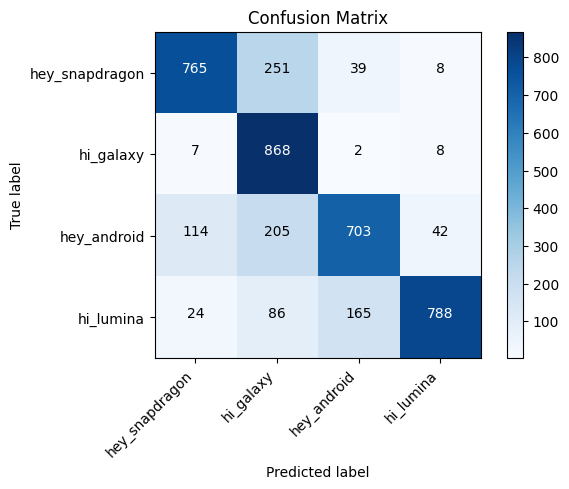


For the student model, accuracy scores are calculated in each stage of the training process for both background excluded dataset and background included dataset and they are compared with each other.

First, a classification model yielded an accuracy of 76.66% following the application of the student model, conducted in the absence of any background noise. A confusion matrix without normalization was generated, depicting actual classes in the rows and predicted classes in the columns for the four keywords—“hey snapdragon,” “hi galaxy,” “hey android,” and “hi lumina”—extracted from the Qualcomm dataset. The matrix revealed an overall reasonable but not highly accurate prediction, given that the model had not been exposed to or trained on the actual data at that point.

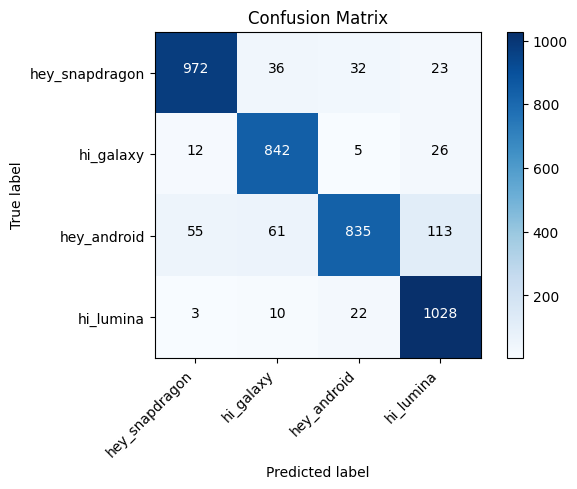
**Figure 2.** *Confusion Matrix for the four keywords: “hey snapdragon”, “hi galaxy”,*

*“hey android”, and “hi lumina” without background and before few-shot DML*



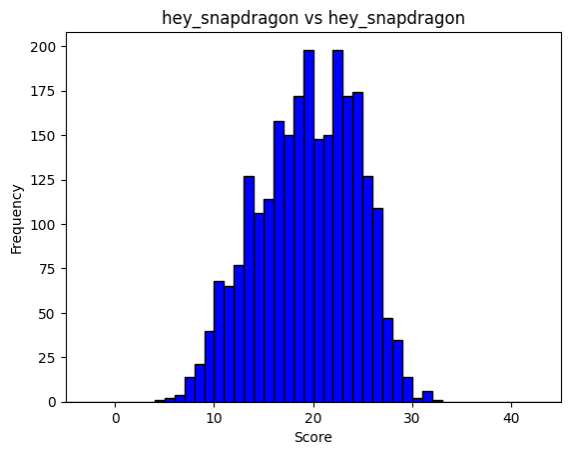
Following the implementation of few-shot Distance Metric Learning (DML), a revised accuracy of 90.23% was computed, and a new confusion matrix was generated post the few-shot DRL. It is evident that exposing the model to just 50 samples per class led to an approximate 15% increase in test accuracy. Notably, this step required less than 2 minutes of training, signifying a cost reduction in both training and deploying models.

**Figure 3.** *Confusion Matrix for the four keywords: “hey snapdragon”, “hi galaxy”,*

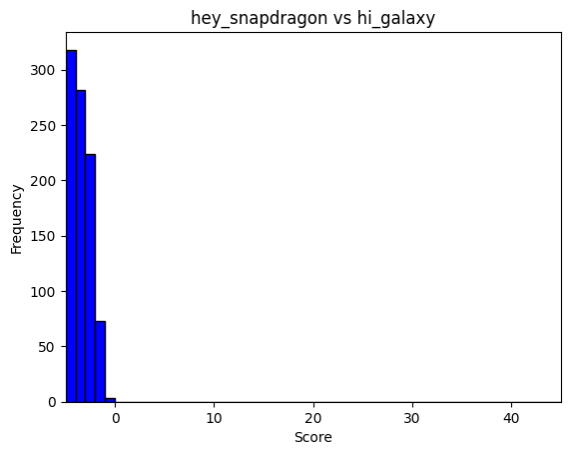
*“hey android”, and “hi lumina” without background and after few-shot DML2*

The distribution of similarity score, which is calculated as the inner product of two embedding vectors, can also show the satisfying result of the training. Take “hey\_snapdragon” as an example as the four keywords all showed a very similar pattern. The similarity scores of “hey\_snapdragon” and itself are all greater than 0, and the median is about 20, while the similarity scores of “hey\_snapdragon” and “hi\_galaxy” are all less than 0. As the higher the inner product, the more similar the two vectors are, this model can discriminate different keywords well. The distributions are plotted and plots for example “hey\_snapdragon” are included.

**Figure 4.** *Distribution of similarity scores between “hey snapdragon” and “hey snapdragon”*



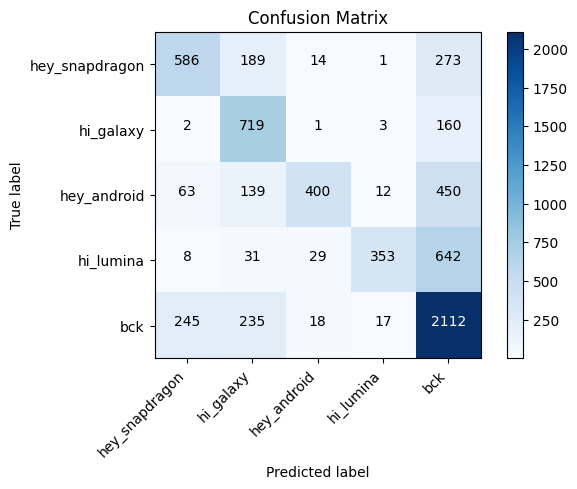
**Figure 5.** *Distribution of similarity scores between “hey snapdragon” and “hi galaxy”*



The background was added into our training data to avoid wrong recognitions. The test accuracy of zero-shot classification with background was 62.22% before applying DML, compared to the baseline accuracy(93.93%) of the teacher model. The superiority of the teacher model can be attributed to its substantial reliance on Wav2Vec2, which is more complex and better-trained than the student model. The embeddings generated by Wav2Vec2 are highly accurate. Even after dimensionality reduction to 128, they still have rich content and the information loss is minor compared to the student model developed and used in this research. In conclusion, the student model still needs improvement.

**Figure 6.** *Confusion Matrix for the four keywords: “hey snapdragon”, “hi galaxy”,*

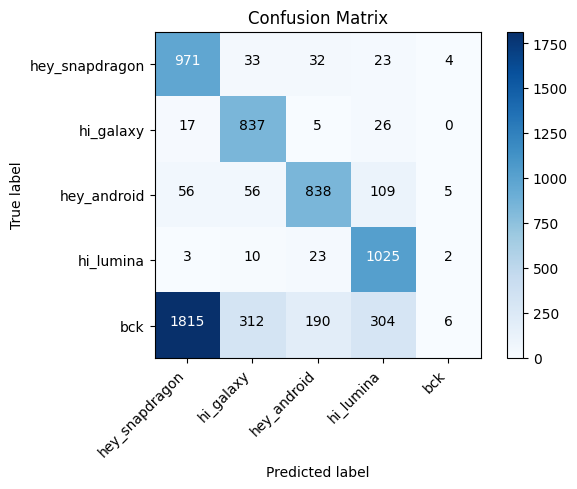
*“hey android”, and “hi lumina” and background before DML*



However, the test accuracy of zero-shot classification was not increased after directly implementing DML but decreased to 54.86%. This time the accuracy after DML and few-shot DRL is lower because the dimensionality reduction caused a certain degree of information loss. The background itself is also complex to analyze and the sample distribution is not balanced.

**Figure 7.** *Confusion Matrix for the four keywords: “hey snapdragon”, “hi galaxy”,*

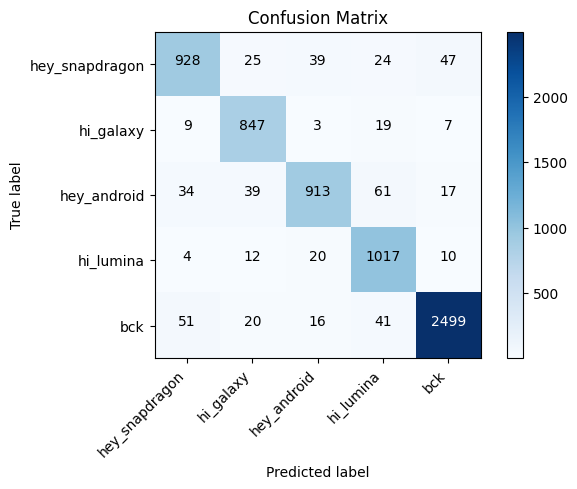
*“hey android”, and “hi lumina” and background after DML*



After training the model for 10000 epochs, an accuracy of 92.57% was achieved with the dataset that includes background from the few-shot classification.

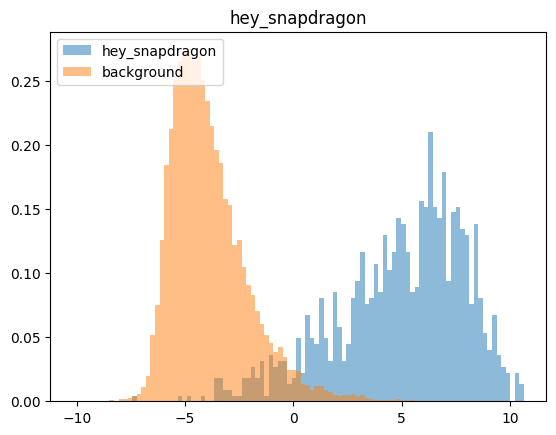
**Figure 8.** *Confusion Matrix for the four keywords: “hey snapdragon”, “hi galaxy”,*

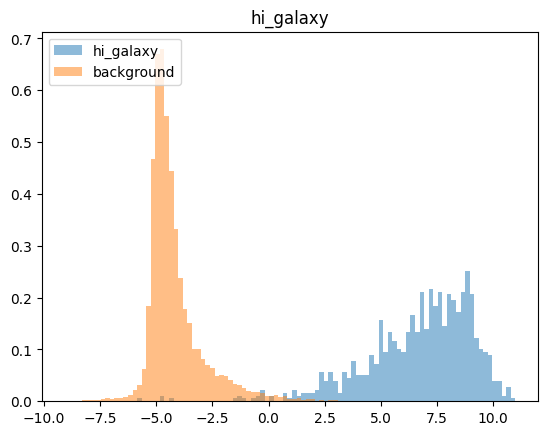
*“hey android”, and “hi lumina” with background after tuned DML*

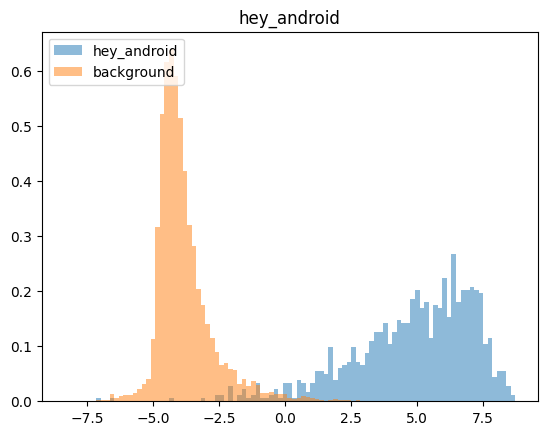


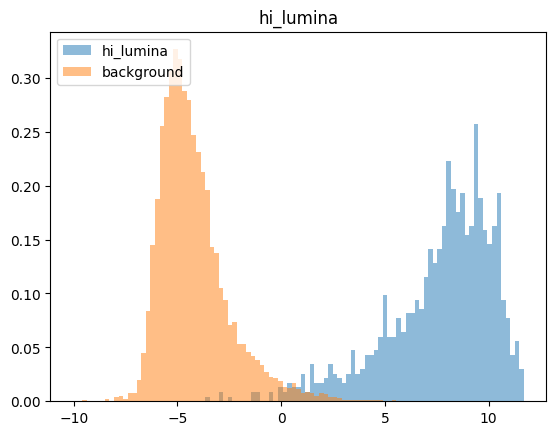
The inner products of embeddings were calculated based on the model after DML was applied, and then the nearest neighbor classification was performed on the training and testing data between each testing class, the keyword, and the background. Based on the nearest neighbor classification, the distribution of nearest neighbor similarity scores for each keyword can be calculated and are plotted.

Histograms of two types of scores: "friends" and "foes", were plotted in the same graph. "Friends" scores represent the similarity between test samples and training samples for the same keyword, while "foes" scores represent the similarity between test samples for the keyword and training samples for the "background" class. Although there is some overlap between two classes, the main part of the keyword 'hey\_snapdragon' is distinguished from the background.

**Figure 9.** *Distribution of similarity score for keyword “hey snapdragon”*

**Figure 10.** *Distribution of similarity score for keyword “hi galaxy”*

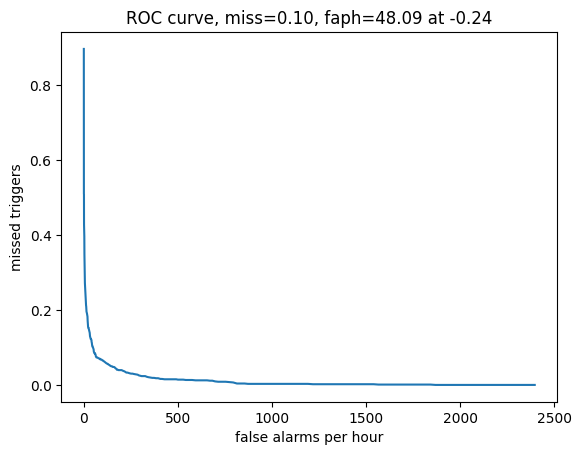
**Figure 11.** *Distribution of similarity score for keyword “hey android”*

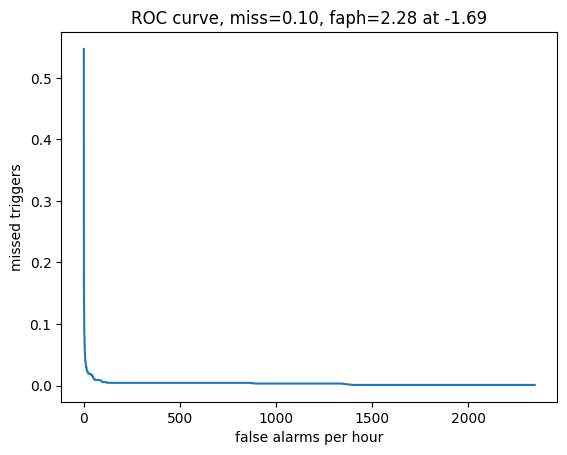
**Figure 12.** *Distribution of similarity score for keyword “hi lumina”*

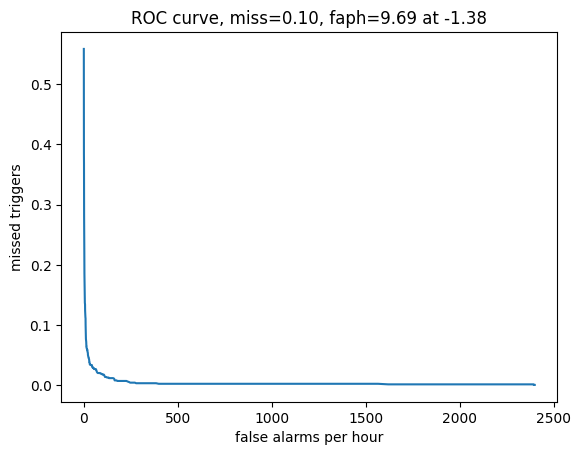
As shown in the distribution figures, there are overlaps between the distribution of similarity scores of keywords and backgrounds. Therefore, it is important to find the thresholds for classification. In order to depict the trade-off between the false negative rate (missed trigger) and the false positive rate (false alarm), a curve that shows the relationship between false alarms and missed triggers of the keywords are plotted. It divides the similarity score range into bins and calculates the false positive rate and miss rate for each bin. The corresponding threshold at 90% recall was also calculated and displayed in each plot as well as the false alarm per hour (FA/h).

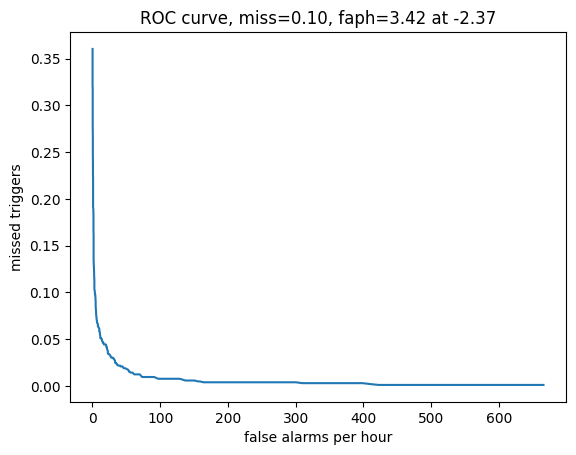
The curve shows how the system's performance varies with different threshold values on the similarity scores. It helps to visualize how well the model can distinguish between the keyword and the background class.

For example, for the keyword 'hey\_snapdragon', in the case of a similarity score -0.01, the recognition rate of the model reaches 90%, and 45.81 false alarms will be generated per hour.

**Figure 13.** *False Alarms vs Missed Triggers curve of keyword “hey snapdragon”*

**Figure 14.** *False Alarms vs Missed Triggers curve for keyword “hi galaxy”*

**Figure 15.** *False Alarms vs Missed Triggers curve for* *keyword “hey android”*

**Figure 16.** *False Alarms vs Missed Triggers curve for keyword “hi lumina”*

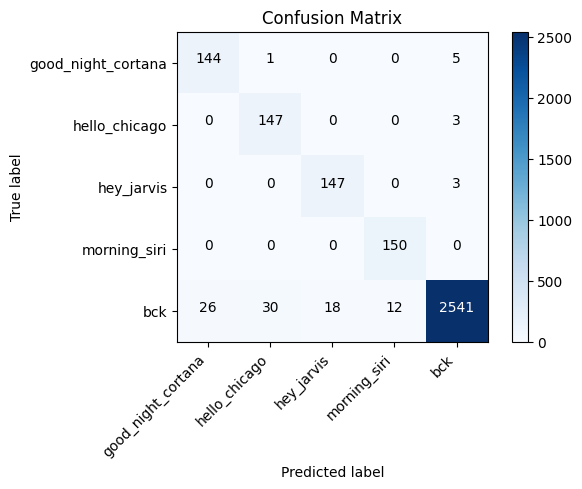
The following is a table showing FA/h for >90% accuracy of all the four keywords. 'Hey\_snapdragon' gets the highest false alarm rate, while 'hi\_galaxy' gets the lowest, which can be speculated that the more complex or longer the keywords, the weaker the model’s recognition ability.

**Table 1.** *False alarm per hour and similarity score of four keywords*

|  | false alarms per hour | threshold similarity |
| --- | --- | --- |
| 'hey\_snapdragon' | 48.09 | -0.24 |
| 'hi\_galaxy' | 2.28 | -1.69 |
| 'hey\_android' | 9.69 | -1.38 |
| 'hi\_lumina' | 3.42 | -2.37 |

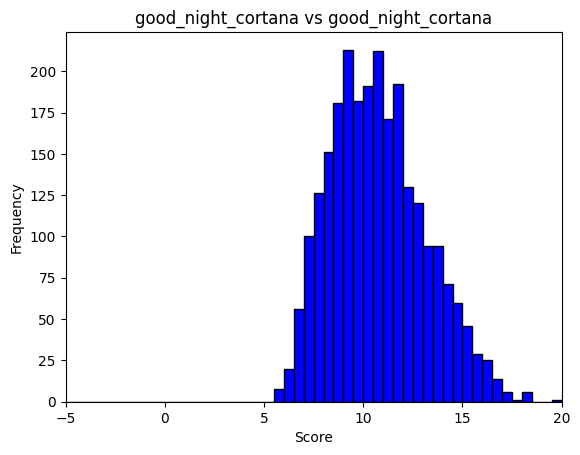
In addition, a simple validation with self-generated Piper dataset was conducted and an overall accuracy of 96.96% was achieved. As intended by the model, only 50 samples from each class were used for training, and the training time was less than 2 minutes. The confusion matrix is included.

**Figure 17.** *Confusion Matrix for the four keywords: “good night cortana”, “hello chicago”, “hey jarvis”, and “morning siri” with background after tuned DML*

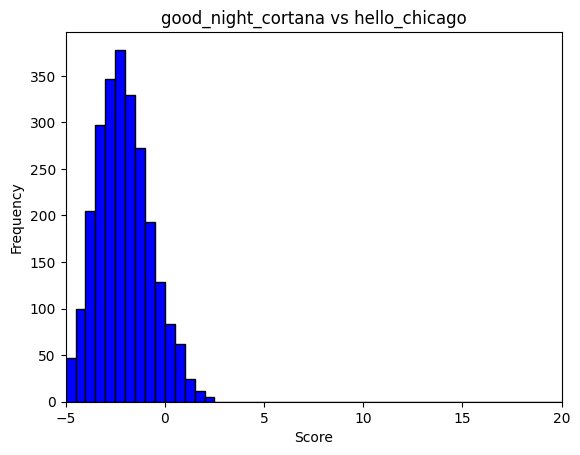
**

Similarly, the distribution of similarity scores for the four new testing keywords are plotted. Take “good night cortana” as an example. The mean and median of similarity scores for “good night cortana” and itself is approximately 12, while if for “good night cortana” and “hello chicago” or with the background is smaller than 0. It shows that the model is able to distinguish between a new and unseen keyword from everyday conversations or other keywords. This pattern is found in all four new keywords.

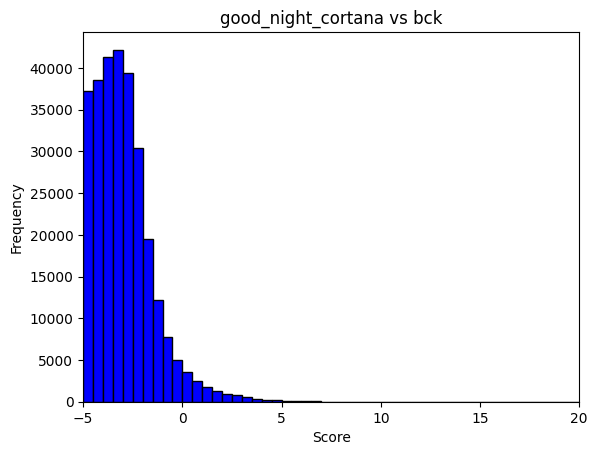
**Figure 18.** *Distribution of similarity scores between “good night cortana” and itself*

**

**Figure 19.** *Distribution of similarity scores between “good night cortana” and “hello chicago”*

**

**Figure 20.** *Distribution of similarity scores between “good night cortana” and background*

**

In conclusion, the findings showed that this teacher-student keyword spotting model is meeting the expectations. First, the dimensionally reduced student model is able to mimic the teacher model and accurately detect the original keywords with an accuracy rate of 92.57% after the few-shot DRL and DML layers compared to the accuracy of 93.93% from the teacher model. Second, the student model is able to achieve this accuracy of over 90% with only 50 samples for training in less than 2 minutes. This training sample requirement and training time is significantly less than traditional deep learning models, making it feasible for customizing. Third, the developed model is able to achieve 96.96% accuracy for self-generated, unseen keywords data. It proved that the model is in fact generalizable to new keywords and is suited for customizable keyword detection tasks.

**Discussion**

The exploration of state-of-the-art acoustic deep learning models presented in this study reveals a significant leap in the performance of audio event identification tasks. While objectives were largely met with a 92.57% accuracy for keyword detection of the four original keywords with background and 96.96% of the new self-generated keywords, the extent of achievement varied, influenced by the inherent constraints of the research.

The model developed for compatibility with the MAX78000 chip showed promising results. It capitalized on the strength of a Teacher-Student Model framework, demonstrating enhanced energy efficiency and processing speed. Despite not achieving all projected metrics, the advancements indicate substantial progress, underscoring the importance of continued refinement.

The methodology that emerged as most effective was the integration of few-shot learning with dimensionality reduction techniques, coupled with DML. This approach yielded an efficient balance between model complexity and performance, crucial for deployment on the specialized hardware. The compact student models, though limited in their capacity compared to the extensive teacher network, managed to perform with notable precision, benefiting from the distilled knowledge.

Alternatives did not reach the same level of effectiveness, primarily due to limitations in handling the complexity of audio data within the constrained parameters of the MAX78000 chip. Some assumptions, such as the scalability of certain models without significant loss in accuracy, were invalidated in practical scenarios. The dimensionality reduction techniques, while beneficial, also introduced a trade-off between information preservation and computational efficiency.

Encountered issues included the non-linearity in audio data, which posed challenges in modeling. Moreover, the balance between model size and accuracy highlighted the difficulty in achieving a model that is both lightweight and highly accurate.

The limitations of the approach are notable. The Teacher-Student Model, while innovative, depends heavily on the quality of the teacher network. The limitations of computational resources also constrained the extent of model complexity that could be achieved. This impacts the outcome by potentially limiting the model's applicability to a broader range of audio events or more challenging acoustic environments.

In conclusion, while the study advances the field of acoustic event detection, particularly within the constraints of ADI’s specialized hardware, it also sheds light on the nuanced challenges of model optimization. The current state of ADI’s technology benefits from these advancements, yet the outcomes signal that the path to an ideal solution is iterative and demands continuous exploration within the boundaries of existing technology and beyond.

Dimensionality posed challenges in data handling. Various dimensionality reduction techniques were employed to address this. Distance metric learning models, specifically designed to minimize the dimensions of embeddings, were implemented. It has been proved that these techniques indeed amplified the efficiency in audio event identification tasks.

Concurrent efforts led to the creation of a model adhering to specific hardware constraints. The model was developed to be compatible with ADI's MAX78000 chip. This initiative was fraught with complexities, such as the requirement to adhere to a confined kernel size and a stipulated number of layers. Notwithstanding these challenges, the emphasis was on achieving energy-efficient operations and ensuring alignment with quantized audio. The result underscores the optimization efforts and their success in enhancing compatibility.

Furthermore, another limitation is that the validation for generalizability was relatively basic. It is performed on only four self-selected keywords with 50 training samples and 150 testing samples for each keyword. The keywords used for validation can only cover a small portion of application cases of actual keywords that may be needed for customization. Improvements can be made by carefully selecting various keywords that cover more real-world scenarios, and generating more data samples for testing to get a fairer accuracy.

Future research seeking solutions of using the teacher-student model for customizable keyword spotting can focus on exploring alternatives to Wav2Vec2 as the foundation for the teacher model. Automatic speech recognition models such as Whisper might be tested and explored. Researchers can also try developing custom teaching models for their specific needs or limitations. Investigating further dimensionality reduction techniques to optimize performance in constrained computational settings may also be a future research option.

**Conclusion**

The teacher-student keyword spotting model with teacher model built based on Facebook Wav2Vec2 successfully achieved the goals of being accurate, efficient, and customizable.

Findings showed that the dimensionally reduced student model trained can successfully emulate the teacher model, accurately identifying each of the four original keywords, "Hey Android," "Hey Snapdragon," "Hi Galaxy," and "Hi Lumina", with an accuracy of 92.57%. It is close to the teacher model's accuracy of 93.93%, which indicates that the performance of the student model is similar to the teacher model and is satisfying.

Additionally, the student model attains a commendable accuracy exceeding 90% with only 50 training samples within a time frame of less than 2 minutes. It is a substantial improvement over the traditional deep learning models, rendering it feasible for customization because the sample size needed is small enough to be collected by individuals and the training time is short enough to be used in real-world applications.

Furthermore, the model demonstrates adaptability and customizability by achieving a 96.96% accuracy in detecting each of the four self-generated, unseen keywords, “good night cortana”, “hello chicago”, “hey jarvis”, and “morning siri”, underscoring its generalizability to unseen keywords and suitability for customizable keyword detection tasks.

The initiative has culminated in a groundbreaking technique for the discernment of audio events, leveraging the Teacher-Student Model paradigm. Here, the “teacher”— a pre-trained, extensive neural network — imparts knowledge to compact 'student' models, streamlining the training process. The ingenuity of this framework is especially advantageous for ADI’s specialized CNN hardware, the MAX78000 and MAX780002 chips.

Compared to existing models, the new framework is engineered to enhance processing velocity, which in turn accelerates speed to market and, consequently, hastens revenue realization. This acceleration is in part attributable to the model's few-shot learning capabilities, which markedly condense training durations while concurrently amplifying accuracy. In addition, the integration of dimensionality reduction with Deep Metric Learning further mitigates computational demands, trimming operational times substantially.

This student model led to a decrease in the collection or preparation effort of large amounts of new customized keyword training audio samples with the minimum of 20 samples. It also reduced the training time on new customized keywords by an estimated 70% and reduced the computational power demand of GPU, saving about $60 for each new customized keyword. This not only positioned ADI to broaden its footprint in the custom audio detection market but also attracted a new clientele. To sum up, this ingenuity is poised to bolster Analog Devices' revenue.

It is prudent to assess the model's prowess in real-world applications to confirm its versatility and operational efficiency across diverse settings. The opportunity to sharpen and diversify the model's capabilities beckons, with potential applications spanning across various sectors of audio and perhaps even extending into video analytics.

In the present landscape, deep learning is indispensable to audio event detection. However, the formidable requirement for extensive training has been a persistent constraint, accentuated when bespoke solutions are sought. The introduced solution meets this challenge squarely, delivering a model that is both efficient and versatile, aligning seamlessly with technical and commercial goals.

Looking ahead, this model could be assimilated into expansive applications such as smart home systems, security frameworks, or healthcare monitoring devices. The platform established here offers vast potential for furtherance, poised to revolutionize the landscape it inhabits.

**References**

Abbasi, S., Hajabdollahi, M., Karimi, N., & Samavi, S. (2019). Modeling Teacher-Student Techniques in Deep Neural Networks for Knowledge Distillation. arXiv preprint arXiv:1912.13179. Retrieved from<https://doi.org/10.48550/arXiv.1912.13179>

Hu, C., Li, X., Liu, D., Chen, X., Wang, J., & Liu, X. (2022). Teacher-Student Architecture for Knowledge Learning: A Survey. arXiv preprint arXiv:2210.17332v1. Retrieved from<https://arxiv.org/abs/2210.17332v1>

Jia, K., Wang, Y., Chen, X., Hu, M., & Tian, Y. (2019). Distributed Learning of Deep Neural Network over Multiple Agents. Journal of Machine Learning, Vol 2, No 1, pp 33-42.

López-Espejo, I., Tan, Z. H., Hansen, J., & Jensen, J. (2021). Deep Spoken Keyword Spotting: An Overview. rXiv preprint arXiv:2111.10592v1 [cs.SD].

Ng, R. W. M., Liu, X., & Swietojanski, P. (2018). Teacher-Student Training for Text-Independent Speaker Recognition. *In 2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1044-1051).

Nguyen, H., Tran, D., & Lu, W. (2020). Teacher-Student Learning with Multiple Student Networks. In Proceedings of the 20th International Conference on Machine Learning (ICML) (pp. 2084-2092).

Park, H. J., Zhu, P., Moreno, I. L., & Subrahmanya, N. (2021). Noisy student-teacher training for robust keyword spotting. rXiv preprint arXiv:2106.01604.

Patrick von Platen. *Wav2Vec2*. <https://huggingface.co/docs/transformers/model_doc/wav2vec2>.

Rhasspy. *GitHub - rhasspy/piper: A fast, local neural text to speech system*. GitHub. <https://github.com/rhasspy/piper>.

Tucker, G., Wu, M., Sun, M., Panchapagesan, S., Fu, G., & Vitaladevuni, S. (2016). "Model compression applied to small-footprint keyword spotting." *In Proceedings of Interspeech 2016* (pp. 2633-2637).

Wang, Y., Wong, J. H. M., Gales, M. J. F., Liu, X., & Ng, R. W. M. (2018). Sequence teacher-student training of acoustic models for automatic free speaking language assessment. *In 2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 463-469).

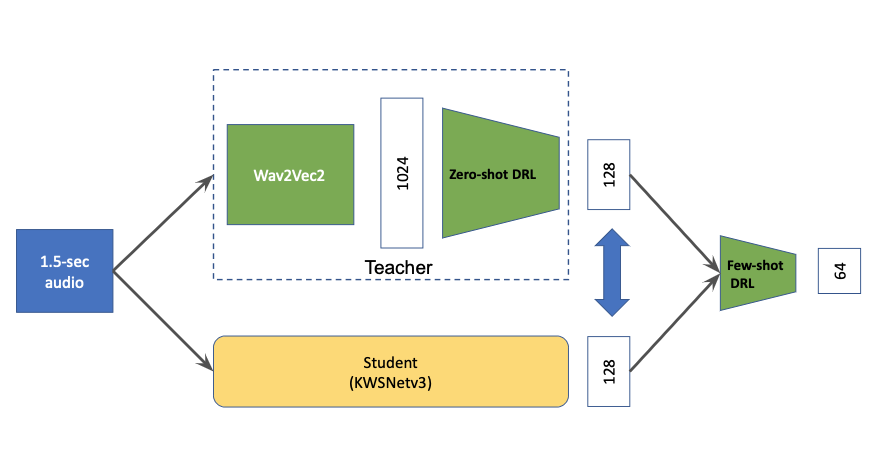
Yao, T., Pan, Y., Ngo, C.W., Li, H., & Mei, T. (2019). Deep Learning for Event-Driven Stock Prediction. In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 2327-2333).

Zhou, Z., Shin, J., Zhang, L., Gurudu, S.R., Gotway, M.B., & Liang, J. (2020). Fine-grained Recurrent Neural Networks for Automatic Prostate Segmentation in Ultrasound Images. In Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (pp. 234-241).

**Appendix A: Model Structure**

The diagram of the basic structure of the teacher-student model is included.

**Figure A1.** *Diagram of the model structure*

****