SMART INDIA HACKATHON 2024

TITLE PAGE

- **Problem Statement ID – 1680**

- **Problem Statement Title-**

  Few Shot Language Agnostic Key Word Spotting system (FSLAKWS) for audio files.

- **Theme- Smart Automation**

- **PS Category- Software**

- **Team ID- 25609**

- **Team Name : PARADOX.**

PARADOX

FALKON

SMART INDIA HACKATHON 2024

Idea Approach [1]

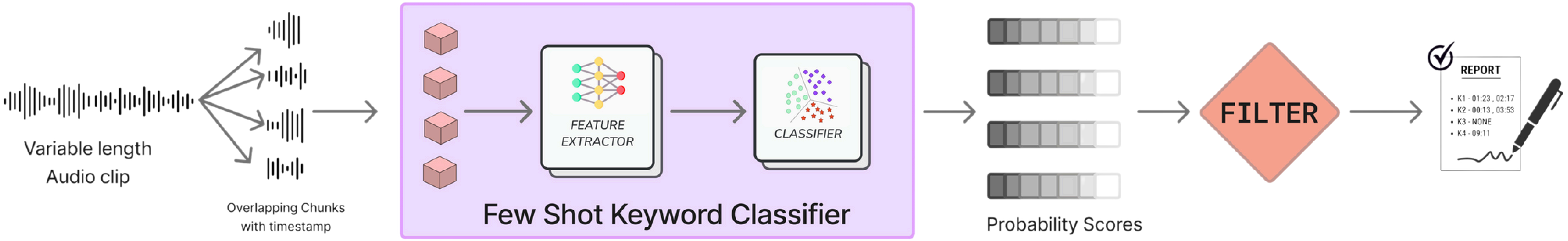
1. We take a variable length audio file as input
2. We resample it to 16kHz
   - **this helps to handle audio files at various sample rates (8 kHz- 48 kHz)**
3. Create overlapping chunks of the audio along with their timestamps (3s each overlayed by 0.5s)
4. Make a 3D tensor representation of the audio by using audio visual representations (Mel spectrogram, Chromagram, Centroid Spectal)
   - **since we are dealing with audio characterstics , hence this make this approach language agnostic**
5. We extract features of these 3D tensors by using transfer learning layers on Resnet further trained on our own data set to provide meaningful 1D features for each audio chunk.
6. Trained a weight generator model to get weights for the new keyword by taking weights of previous keywords and features of few examples of new keyword
   - **Few-Shot Continual Learning [3]**
7. Used cosine similarity and SoftMax function to classify each audio chunk.
8. Generate a report by checking probability scores of each sample and giving timestamp only when threshold is met

Novelty / Uniqueness

- **Data Set Creation [4] :** The dataset was created with multilingual, accented, and diverse audio samples by our own team, and was further augmented using pitch shifting, time stretching, and background noise addition to increase its generalizability.
- **Flexible Feature Extractor :** Any FE(PANN or a MobileNet) can be used inplace of Resnet. The choice can be made based on requirements of accuracy or size.
- **Transfer Learning with ResNet:** Pre-trained weights can be adapted effectively for audio feature extraction reducing the need of large audio dataset.
- **Unique Weight Generator for Keywords [3] :** A dedicated weight generator is used for keyword addition, based on few-shot examples.

Addressing Key Feature Requirements

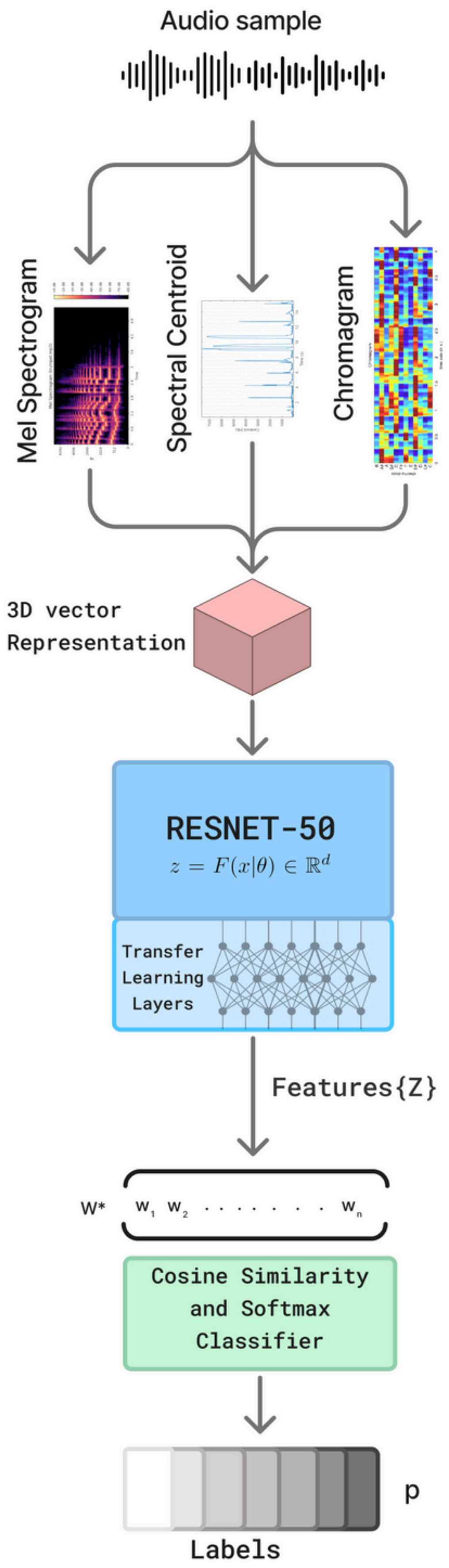
- **Varying Sampling Rate**: Handling varying sample rate by resampling to 16KHz. It won't effect the performance as audio representations has been used.
- **Audio Feature Agnosticism**: Audio representations like Mel spectrogram, Chromagram, Spectral Centroid has been used, hence it becomes language agnostic.
- **Few-Shot Continual Learning [3]** : Trained a weight generator model to get weights for the new keyword by taking weights of previous keywords and features of few examples of new keyword



PARADOX

TECHNICAL APPROACH

SMART INDIA HACKATHON 2024



Feature Extraction and Classification[2]

1.) Preprocessing: The audio is resampled to a target sample rate (16 kHz), the clips were padded with zeros if they were shorter than the required length or truncated if they exceeded the maximum length. This ensured that each input had a uniform size of 48,000 samples (16kHz sample rate × 3 seconds) making the input consistent for the neural network.

2.) Stacking: The features are stacked along the depth dimension (i.e., channel-wise) to create a 3D input of size (128 × 94 × 3) per audio sample for the CNN. For each audio sample, the extracted features (Mel, Spectral Centroid, and Chromagram) are concatenated, resulting in 3 channels for the input, similar to RGB channels in images.

3.) ResNet-50 Transfer Learning: We load a pre-trained ResNet model (on ImageNet), using it as a feature extractor We utilized pre-trained weights for the lower layers, which capture basic features such as edges and textures. The early layers are frozen to retain the knowledge learned from large-scale image data, and the last few layers are fine-tuned for the specific audio classification task.

4.) Dimensionality Reduction We flatten the high dimensional feature map we got as output, to a 1D vector of size 256, which represents a compressed but captures both time and frequency dependencies in the audio sample

5.) Classification: It is done by cosine similarity followed by a SoftMax function. This assigns probabilities to each keyword based on its proximity to the learned feature vectors.

- Cosine similarity measures the similarity between the new audio sample and base keywords.
- SoftMax function was applied to convert these cosine similarity scores into probabilities.

Few-Shot Classification Weight Generator (FSCWG)[3]

- The Few-Shot Classification Weight Generator (FSCWG) is a model designed to create classification weight vector for new category using only a few examples.
- It works by learning from weights of pre-trained base categories and applies that knowledge to generate weights of novel category, even with very limited data.
- The FSCWG is trained by sampling "fake" novel categories from the base categories. For each sampled category, it takes their feature vectors, and generates a weight vector for the "fake" novel category and loss is computed by comparing the original weight with generated weight for the "fake" novel category.
- This weight vector is calculated through two methods:
  1. Feature Averaging: A simple average of the feature vectors (z) of the few examples.

$$w_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} z_i$$

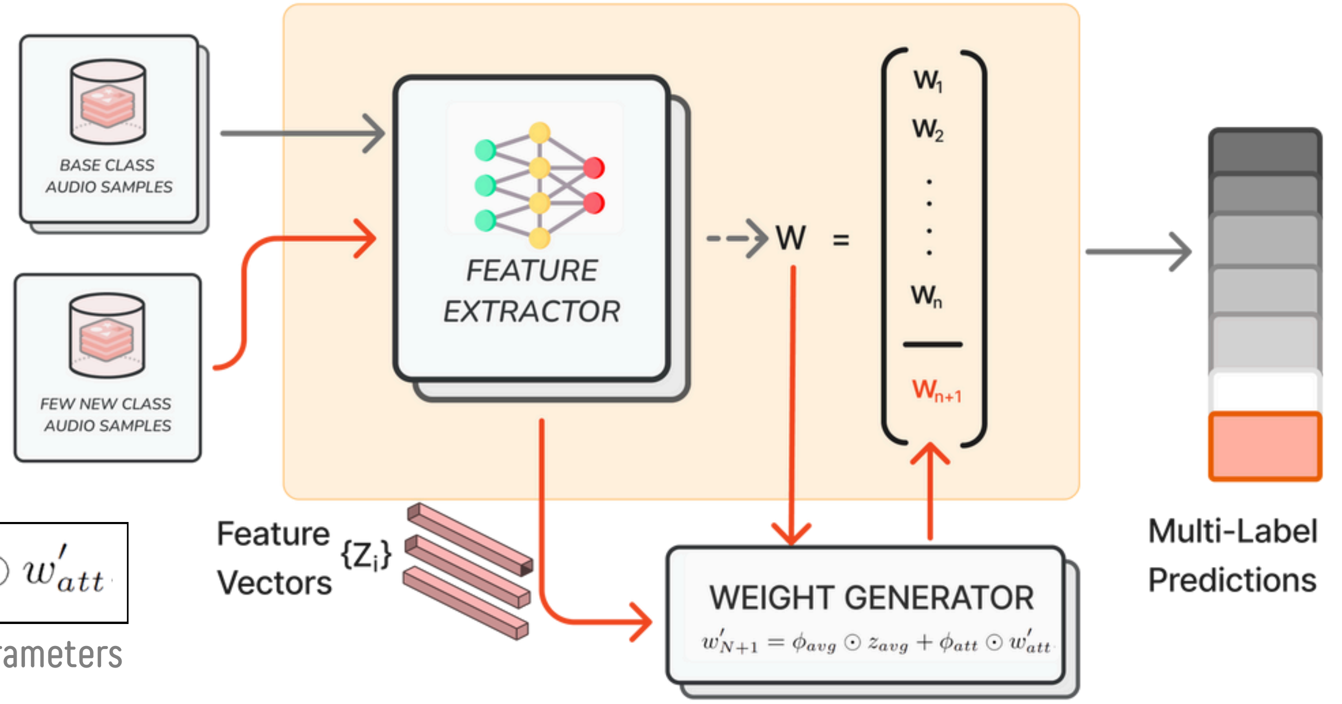
  2. Attention Mechanism: An advanced process where the model assigns scores to the pre-trained weight vectors of the base categories, determining how much they should influence the final weight vector.

$$w_{\text{att}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{b=1}^{K_{\text{base}}} \text{Att}(qz_i, k_b)w_b$$

- The final weight for a new category is a combination of both these methods. The output of the weight generator is a classification weight vector of the novel category.
- The FSCWG enables effective classification of new categories with minimal data by learning from pre-existing knowledge and using attention to enhance accuracy.

$$w'_{N+1} = \phi_{avg} \odot z_{avg} + \phi_{att} \odot w'_{att}$$

ϕ_{avg} and ϕ_{att} are learnable parameters



Tech-stack

- **Librosa:** For audio feature extraction (Mel Spectrogram, Spectral Centroid, Chromagram).
- **NumPy:** For numerical operations, padding, and array manipulations
- **TensorFlow and Keras:** For building and training the deep learning model.
- **Scikit-learn:** For data preprocessing, encoding labels, and train-test splitting
- **ResNet50:** A pre-trained Convolutional Neural Network (CNN) model for transfer learning.

FEASIBILITY AND VIABILITY

FEASIBILITY & VIABILITY

Proven Technology

Our system leverages well-established techniques which are widely used in audio processing. These are language-independent and have been **proven effective in multiple research applications.**

Efficient Architecture

The system includes a small weight generator to **bypass the need for repeated training** of Feature Extractor, optimizing both computation time and resource usage. This ensures the system remains scalable as the dataset grows.

Market Demand

As the world becomes increasingly connected, there is a **growing need** for language-agnostic voice technology, especially in voice assistants, smart devices, and customer support.

LIMITATIONS

Short audio clips

Chunks limited to 3-second clips, making it difficult to process lengthy keywords.

Substring issues

False positive when a substring of a keyword is present. (Cock in Peacock)

Homophones

Similar-sounding words (e.g., "ate" and "eight") cause confusion.

Lack of context

Cannot distinguish between different meanings of the same word, such as "apple" (fruit vs. company).

CHALLENGE	SOLUTION
Language Agnostic	Overcame language barriers by using audio features like mel spectrogram, chromagram, and spectral centroid.
Versatile Dataset	Created a dataset ourselves that was further augmented with diverse accents, multiple multilingual keywords, by applying data augmentation techniques.[4]
Repeated Training	Introduced a small weight generator to eliminate the need to retrain the feature extractor with each new dataset.
Forgetting Previous Knowledge	Implemented regularization techniques and an attention kernel to retain past learnings.

IMPACT AND BENEFITS

BENEFITS

Adaptive Architecture

Feature extraction can **adapt to any model,** such as PANN or MobileNet, with the same flow applied regardless of the choice

Scalability

The model supports online learning, adapting to new keywords while retaining prior knowledge **without needing frequent retraining**, thanks to its attention-based framework.

Flexibility in Audio Formats

The system handles **variable audio formats (8k-48k sample rates)**, ensuring compatibility across devices and environments, especially in industries like telecommunications and media.

Cost Effective & Resource Efficient

FSLAKWS reduces costs by **minimizing the need for large datasets**, making classification feasible even with sparse data.

Social Inclusiveness

Our model trained on audio features, serves diverse populations and promotes inclusivity. FSLAKWS also **supports low-resource languages**, breaking down barriers and enhancing accessibility.

IMPACTS

IOT devices

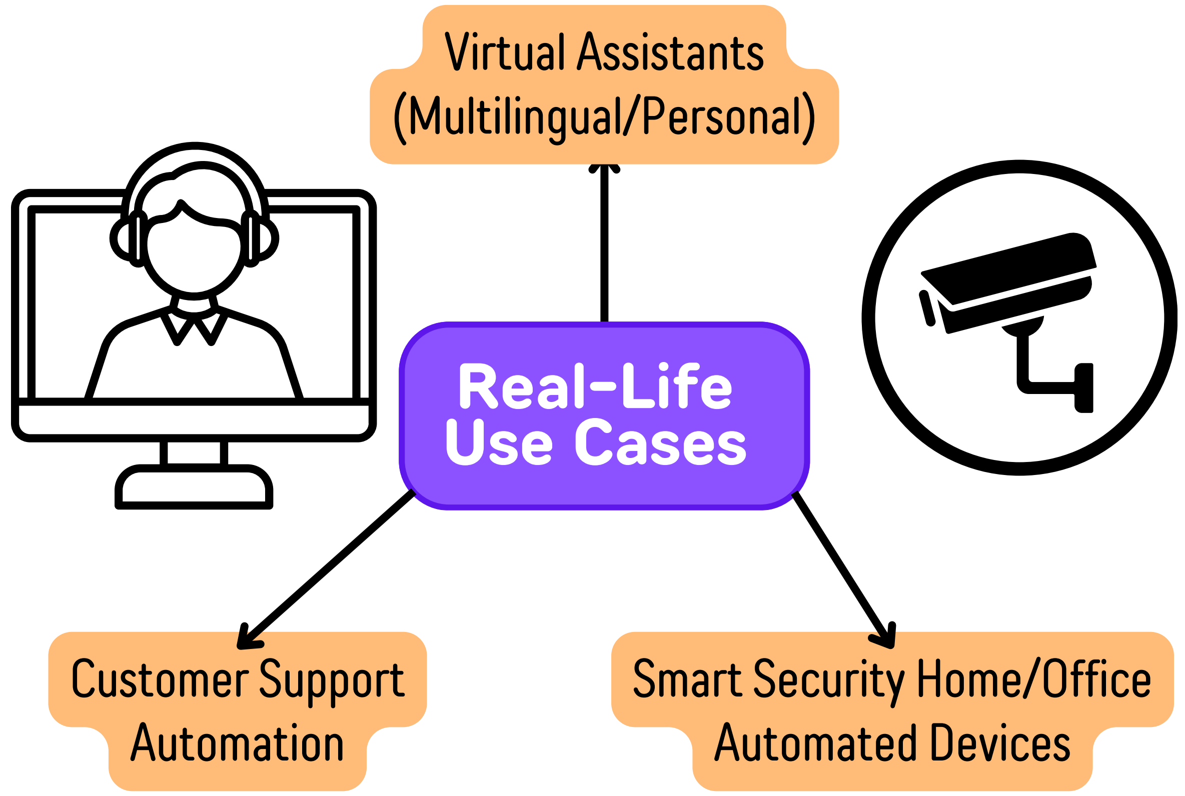
Our architecture can be **applied to any smart device** , providing them with Keyword Spotting capabilities allowing for better communication between User and Device. In IoT ecosystems, audio files recorded from various devices (with different sample rates) can still be processed reliably without additional standardization.

Personalized content delivery

By **analyzing the keywords used by individuals** when searching for audio content, FSLAKWS can enable media companies to deliver highly personalized news and content recommendations, increasing user engagement and satisfaction.

Security and Surveillance

In security contexts, the system can be used for voice command recognition or **audio surveillance across different environments** without being hindered by language barriers.



Real-Life Use Cases

Virtual Assistants (Multilingual/Personal)

Customer Support Automation

Smart Security Home/Office Automated Devices

PARADOX

RESEARCH AND REFERENCES

REFERENCES

[1] IDEA_APPROACH_Detailed_Explaination

[2] TECHNICAL_APPROACH_Detailed_Explaination

[3] WEIGHT_GENERATOR_Detailed_Explaination

[4] DATASET_hugging_face

RESEARCH

FEW-SHOT CONTINUAL LEARNING FOR AUDIO CLASSIFICATION by Yu Wang, Nicholas J. Bryan, Mark Cartwright, Juan Pablo Bello, Justin Salamon

Dynamic Few-Shot Visual Learning without Forgetting by Spyros Gidaris and Nikos Komodakis

CNNs for Audio Classification

YOUTUBE VIDEO LINK

SIH-1680 | FSLAKWS | PARADOX

SMART INDIA HACKATHON 2024-

TITLE PAGE

- Problem Statement ID – 1680
- Problem Statement Title- Few Shot Language Agnostic Key Word Spotting system (FSLAKWS) for a les.
- Theme- Smart Automation
- PS Category- Software
- Team ID- 25609
- Team Name : PARADOX.