A REPORT

On

ASSIGNMENT 1(Speech Understanding)

Shripad pate(M23MAC007)

Data and Computational Sciences (DCS)
Department of Mathematics
Indian Institute of Technology, Jodhpur

Feb 2025

Question 1:

1)Speech Emotion Recognition (SER): Importance and Real-World Applications:

Speech Emotion Recognition (SER) is a specialized subfield of speech recognition that focuses on analyzing speech signals to determine the emotional state of a speaker. Unlike traditional speech recognition, which transcribes spoken language into text, SER aims to extract affective information by analyzing acoustic features such as pitch, intensity, and rhythm. The goal is to classify emotions either as categorical states (e.g., happiness, sadness, anger) or as continuous dimensions like arousal (energy level), valence (positivity/negativity), and dominance (control over the situation).

Types of Speech Recognition Tasks Mentioned in the Paper

The paper discusses different aspects of speech recognition, particularly in the context of Speech Emotion Recognition (SER). Here's how it fits into the broader speech recognition landscape:

- **Automatic Speech Recognition (ASR)**: Converts spoken language into text, forming the basis of technologies like voice assistants (Siri, Alexa), automated transcription services, and voice search.
- **Speech Emotion Recognition (SER)**: Goes beyond transcription by analyzing the speaker's emotional state. This is useful for AI-driven human-computer interaction, mental health monitoring, and customer service.
- **Dimensional Speech Emotion Recognition (A/D/V)**: Instead of classifying emotions into predefined categories, this approach maps emotions onto three continuous dimensions:
    - **Arousal (A) –** Measures energy and intensity (e.g., calm vs. excited).
    - **Dominance (D) –** Represents the level of control or assertiveness (e.g., passive vs. dominant).
    - **Valence (V) –** Determines the emotional polarity (positive vs. negative).

This dimensional approach allows for more nuanced and context-aware emotion detection, making it preferable for real-world applications where emotions are complex and variable.

Real-World Applications of Speech Emotion Recognition (SER)

SER has a broad range of applications across multiple industries, making it a valuable technology for improving human-computer interaction and enhancing emotional intelligence in AI systems. Below are some key domains where SER plays a crucial role:

(a) Healthcare and Mental Health Monitoring

SER is increasingly used in healthcare, especially for diagnosing and monitoring mental health conditions.

- **Depression & Anxiety Detection**: Changes in vocal tone, pitch, and speech patterns can indicate mental health issues. SER can assist in early diagnosis by detecting emotional distress in a patient's voice.
- **Stress Monitoring**: SER can help assess stress levels in high-pressure professions (e.g., air traffic controllers, military personnel) by analyzing speech patterns.
- **Therapeutic Support**: AI-driven therapy bots can use SER to respond empathetically to users, making mental health resources more accessible.

(b) Customer Service & Call Centers

SER is transforming customer interactions by enabling AI-powered systems to understand and respond to customer emotions.

- **Sentiment Analysis in Calls**: Call center analytics can detect frustration or dissatisfaction in customer voices and escalate issues to human agents.
- **Emotion-Adaptive Chatbots**: AI-powered customer service bots can adjust their tone and responses based on the caller's emotions, improving user experience.
- **Quality Assurance**: SER can analyze agent-customer interactions to assess emotional intelligence and service quality.

(c) Human-Computer Interaction (HCI)

Emotion-aware AI assistants and interactive systems benefit from SER, making digital interactions more human-like.

- **Voice Assistants (Siri, Alexa, Google Assistant)**: Adding emotional awareness to AI assistants allows them to respond more naturally, improving user engagement.
- **Gaming & Virtual Reality (VR)**: SER can be used in gaming to adjust in-game experiences based on player emotions. For instance, horror games could adjust difficulty if a player's voice indicates stress.
- **Adaptive Learning Systems**: Online education platforms can detect students' emotional states and adjust lesson pacing accordingly.

(d) Security & Surveillance

SER plays a critical role in enhancing security by detecting stress, fear, or deception in voice recordings.

- **Emergency Call Analysis**: SER can help emergency services prioritize distress calls based on detected emotional urgency.
- **Lie Detection**: Law enforcement agencies use SER to analyze speech patterns for signs of deception in investigations and interrogations.
- **Surveillance & Threat Detection**: SER can assist in monitoring public spaces by identifying individuals exhibiting high-stress or aggressive speech patterns.

(e) Low-Resource Applications & Edge Computing

Deploying SER models on low-power devices is crucial for expanding accessibility.

- **Smartphones & IoT Devices**: SER-enabled apps can provide real-time emotional insights without needing cloud processing.
- **Wearables & Smartwatches**: Health-tracking devices can use SER to monitor stress levels and mental well-being.
- **Car Infotainment Systems**: Modern vehicles can integrate SER to assess driver emotions, issuing alerts if fatigue or frustration is detected.

Despite its potential, SER faces several challenges that researchers are actively working to overcome:

- **Annotator Disagreement**: Emotions are subjective, making it difficult to obtain consistently labeled training data.
- **Data Scarcity**: High-quality, emotion-labeled speech datasets are limited, particularly for underrepresented languages.
- **Cross-Cultural Variability**: Emotional expression varies across cultures, requiring SER models to be adaptable.
- **Computational Constraints**: Deploying SER on low-power devices requires lightweight models like Wav2Small, which balance efficiency and accuracy.

2)Analyze the strengths and limitations of state of the art models or tools in terms of the methods or models available.

Model	Parameters	Memory Usage
Teacher Model (WavLM + Dawn)	483.9M	1929MB
WavLM (2024 Challenge Winner)	318.6M	1284MB
Dawn (Wav2Vec2-based)	165.3M	697MB
MobileNetV4-L	31.87M	257MB
MobileNetV4-M	10.38M	94MB
MobileNetV4-S	3.12M	36MB
Wav2Small	72K	9MB

Analysis of State-of-the-Art Models for Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) holds significant potential but faces multiple challenges that researchers are actively working to address. Key issues include:

- Annotator Disagreement: Emotion perception is inherently subjective, making it difficult to obtain consistently labeled training data.
- Data Scarcity: High-quality, emotion-labeled speech datasets are limited, particularly for underrepresented languages.
- Cross-Cultural Variability: Emotional expressions differ across cultures, requiring SER models to generalize effectively.
- Computational Constraints: Deploying SER on low-power devices necessitates lightweight models that balance efficiency and accuracy.

Comparison of SER Models

Wav2Vec2 / WavLM

Strengths:

- Utilize transformer layers and VGG7 feature extractors.
- Achieve high Concordance Correlation Coefficient (CCC), with valence CCC reaching 0.64.
- Leverage linguistic cues for valence prediction.

Limitations:

- High computational footprint (e.g., 87.9M parameters for Wav2Vec2).

- Slow inference speed, making deployment on low-resource hardware challenging.

MobileNet Variants

Strengths:

- Designed for efficiency (e.g., MobileNetV4-S has only 3.12M parameters).
- Achieve fast execution times (5–11 ms latency).

Limitations:

- Reduced time resolution (16 tokens/s vs. Wav2Small's 250 tokens/s).
- Lower CCC for valence (0.42) compared to transformer-based models.

Proposed Wav2Small

Strengths:

- Ultra-lightweight architecture with only 72K parameters and a memory footprint of 9 MB RAM.
- Achieves competitive performance in arousal prediction (CCC = 0.66 on MSP Podcast dataset).
- Utilizes non-contiguous memory reshaping to preserve token resolution.

Limitations:

- Lower valence CCC (0.37), trailing behind MobileNetV4-S.

Implications and Future Directions

- High-Performance Models: Transformer-based models such as WavLM and Dawn achieve the highest accuracy but demand substantial computational resources.
- Efficiency-Oriented Models: MobileNet-based architectures offer lightweight alternatives but sacrifice some accuracy.
- Wav2Small Innovation: Wav2Small, leveraging knowledge distillation, significantly reduces model size while maintaining competitive performance. By distilling Wav2Vec2 into a 72K-parameter model, Wav2Small offers a practical solution for SER applications in resource-constrained environments.

The development of efficient SER models like Wav2Small represents a promising step toward making emotion recognition more accessible across diverse applications, including mobile devices and edge computing. Future research should focus on improving valence prediction and adapting models for multilingual and culturally diverse datasets.

3)Discuss the results in terms of the metrics used to evaluate the task, including their strengths and limitations.

SER models are typically evaluated using:

- **Concordance Correlation Coefficient (CCC):** Measures agreement between predicted and ground truth values.
  - **Strength:** More robust than L2 distance in A/D/V tasks.
  - **Limitation:** Does not always reflect extreme values well.
- **Quadrant Correction Loss (Newly Proposed):** Penalizes quadrant mismatches between student and teacher models.
  - **Strength:** Helps align predictions with human perception.
  - **Limitation:** May not fully capture subtle emotion variations.

The **CCC results on IEMOCAP** dataset confirm that Wav2Small performs comparably to MobileNets while using fewer parameters.

Key Results:

- **Teacher Model** (WavLM + Wav2Vec2 ensemble): Sets SOTA valence CCC = 0.676 on MSP Podcast.
- **Wav2Small**: Achieves arousal CCC = 0.66 (MSP Podcast) and 0.56 (IEMOCAP), demonstrating cross-dataset robustness.

4)Suggest what are the open problems and opportunities corresponding to that problem statement.

Key Challenges:

1. **Improving Valence Recognition**: Predicting valence (positive/negative emotion) remains more challenging than arousal and dominance due to its subjective nature and subtle variations in speech.
2. **Extreme Value Prediction**: Current state-of-the-art (SotA) models tend to avoid extreme predictions, leading to conservative outputs that do not fully capture the emotional spectrum.

3. **Fusion of Text and Audio**: Incorporating linguistic cues can improve recognition accuracy but introduces risks such as accent bias and language dependence.
4. **Dataset Distillation**: The generation of synthetic datasets using teacher models could improve training efficiency and model performance.

Potential Research Directions:

1. **Neural Architecture Search (NAS)**: Automating model discovery to optimize SER architectures can lead to more efficient and accurate systems.
2. **Multi-modal Emotion Recognition**: Integrating text, audio, and facial cues to enhance emotion recognition capabilities.
3. **Robustness Testing**: Evaluating SER models in noisy environments and cross-lingual settings to improve generalization and fairness.

Open Problems and Opportunities:

1. **Extreme Value Prediction:**
   - *Problem:* Label averaging in datasets leads to conservative predictions, reducing sensitivity to extreme emotions.
   - *Opportunity:* Developing novel loss functions or augmentation techniques to encourage extreme value recognition while maintaining reliability.
2. **Noise and Language Fairness:**
   - *Problem:* The Wav2Small model passes 74% of noise and language fairness tests compared to 79% for the teacher model, highlighting robustness issues.
   - *Opportunity:* Improving domain adaptation techniques and leveraging adversarial training to enhance model robustness against environmental variations and linguistic diversity.
3. **Data Scarcity:**
   - *Problem:* There is a lack of diverse and representative datasets, particularly for arousal, dominance, and valence (A/D/V) in natural settings (e.g., MSP Podcast dataset is rare).
   - *Opportunity:* Expanding existing datasets, creating synthetic data through generative models, and leveraging transfer learning from related tasks.
4. **Linguistic Bias:**
   - *Problem:* Over-reliance on linguistic cues for valence recognition can introduce accent and language bias.
   - *Opportunity:* Investigating language-independent emotion features and exploring self-supervised learning techniques to reduce bias.

Conclusion:

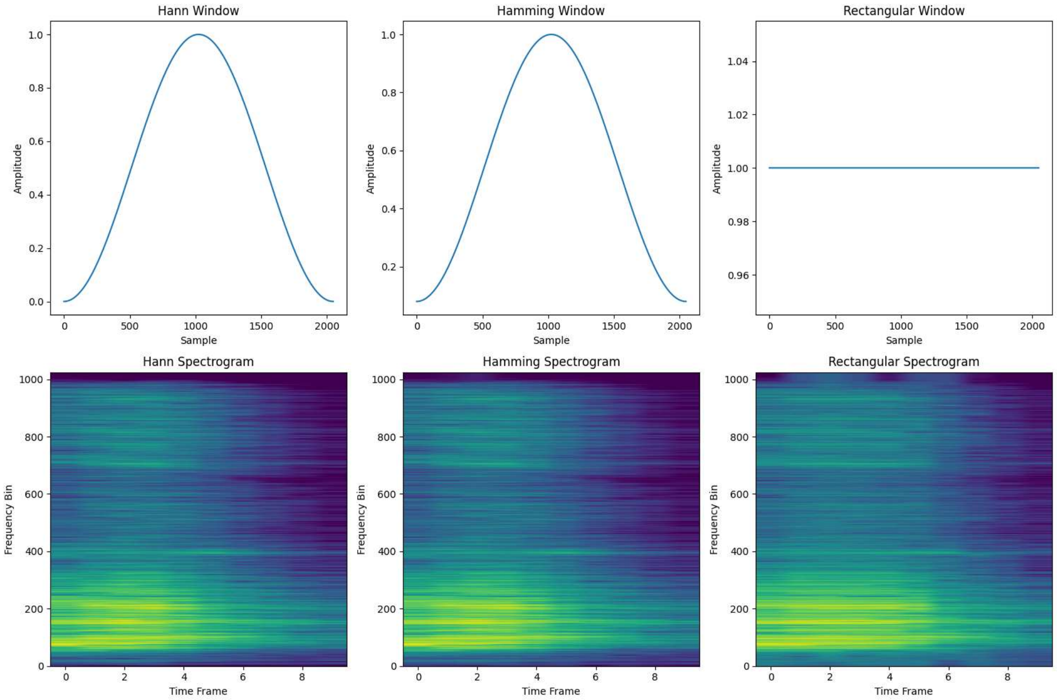
The paper presents an effective approach to reducing the size of large speech recognition models while maintaining strong performance in speech emotion recognition. The knowledge distillation technique, transferring learned representations from Wav2Vec2 to the smaller Wav2Small model, demonstrates promising results on evaluation datasets. However, the approach has primarily been tested on SER tasks, leaving questions about its broader applicability to other speech-related domains such as automatic speech recognition (ASR) and speaker identification. Additionally, further details on the knowledge distillation process and architectural choices would provide deeper insights into the key factors contributing to the model's efficiency and performance.

Overall, Wav2Small represents a significant step toward making large speech models more accessible for real-world, low-resource applications. Future research should focus on enhancing model robustness, mitigating bias, and expanding the generalizability of the proposed approach to other speech-related tasks.

Question 2:

Task1:

In the task 1 we are apply three different windowing technique and analysis the results here we apply three windowing in spectrogram done by STFT



Windowing Techniques:

- **Hann Window:**
  - The Hann window tapers smoothly to zero at both ends.
  - It reduces spectral leakage significantly, resulting in a smoother spectrum in the spectrogram.
- **Hamming Window:**
  - The Hamming window also tapers but retains a small value at the ends.
  - It provides a compromise between resolution and spectral leakage.
- **Rectangular Window:**
  - The amplitude is constant across the window.
  - This causes higher spectral leakage and less smoothing, making frequency components harder to distinguish.

2. Spectrogram Analysis:

- **Hann Spectrogram:**
  - The transitions and frequency components are smoother and more distinct due to reduced spectral leakage.
  - Preferred when smoothness in frequency representation is required.
- **Hamming Spectrogram:**
  - Similar to the Hann window but with slightly more pronounced spectral leakage.

    - Offers a balance between resolution and leakage.
- **Rectangular Spectrogram:**
    - Exhibits higher spectral leakage, leading to blurred transitions and less clarity in frequency components.
    - While simpler computationally, it is less effective for frequency analysis.

3. Comparison:

- Both the Hann and Hamming windows produce spectrograms with better-defined frequency bands compared to the rectangular window.
- The rectangular window lacks the tapering effect, making it less ideal for applications requiring detailed frequency analysis.
- Hann offers the best leakage reduction, making it suitable for applications like speech or music analysis, while Hamming can be used where a trade-off is acceptable.

Conclusion:

The choice of the windowing technique significantly impacts the clarity of the spectrogram. Based on these spectrograms:

- **Hann window** provides the best balance for minimizing leakage and preserving spectral details.
- **Hamming window** is a good alternative with similar but slightly less pronounced benefits.
- **Rectangular window** is less effective for resolving frequency details and is not recommended for complex signal analysis.

UrbanSound8K Dataset and Training Procedure

1. Overview of UrbanSound8K Dataset

The UrbanSound8K dataset is a well-known dataset used for environmental sound classification. It contains 8,732 labeled audio samples categorized into 10 different sound classes, including:

1. Air Conditioner
2. Car Horn
3. Children Playing
4. Dog Bark
5. Drilling
6. Engine Idling

7. Gun Shot
8. Jackhammer
9. Siren
10. Street Music

Each sound clip in the dataset is ≤ 4 seconds long and is sampled at 44.1 kHz. The dataset is split into 10 folds, enabling cross-validation techniques.

2. **Preprocessing and Feature Extraction**

Since the model is based on ResNet50, which is designed for image classification, we convert the raw audio signals into spectrograms using Short-Time Fourier Transform (STFT) with three different windowing techniques:

- Hann Window
- Hamming Window
- Rectangular Window

Each windowing method affects the frequency resolution of the generated spectrograms, impacting model performance.

Steps in Preprocessing:

1. Load the raw audio files.
2. Apply STFT to compute the spectrograms.
3. Normalize spectrogram values to improve model generalization.
4. Convert them into grayscale images (1 channel) to match the ResNet input format.
5. Resize the images to 224×224 to fit ResNet50.

---

3. Model Architecture - ResNet50 as Backbone and Training Procedure

The ResNet50 architecture is a deep convolutional neural network (CNN) with residual connections, allowing it to train effectively without vanishing gradient issues.

The training process follows these steps:

Step 1: Data Splitting

- The dataset is split into training (80%) and validation (20%).
- The PyTorch DataLoader is used to batch and shuffle data for training.

Step 2: Loss Function and Optimizer

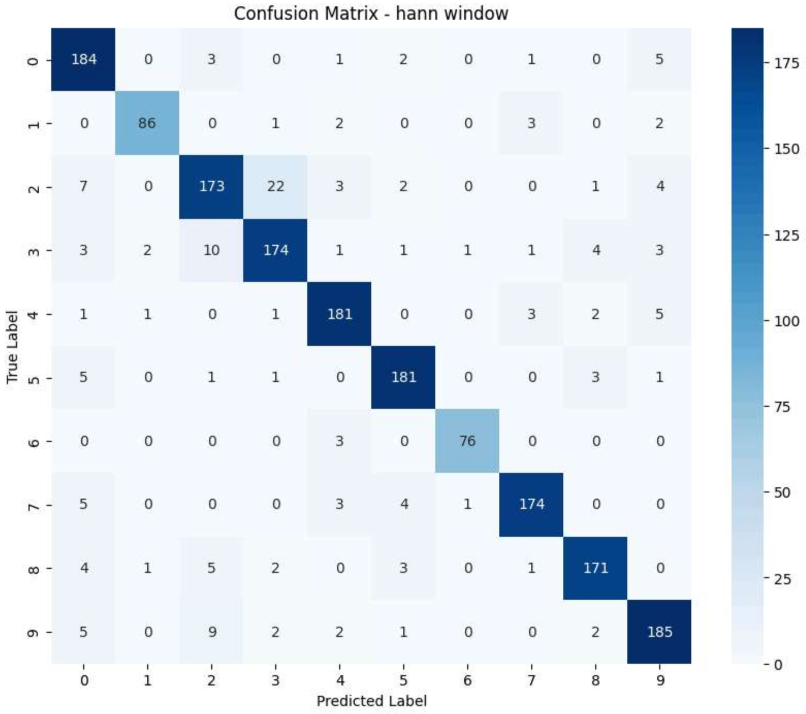
- Loss Function:
  - We use CrossEntropyLoss, suitable for multi-class classification.
- Optimizer:
  - AdamW optimizer with L2 weight decay for regularization:
- Learning Rate Scheduler:
  - OneCycleLR adjusts the learning rate dynamically, with an initial warm-up phase

4. Evaluation Metrics

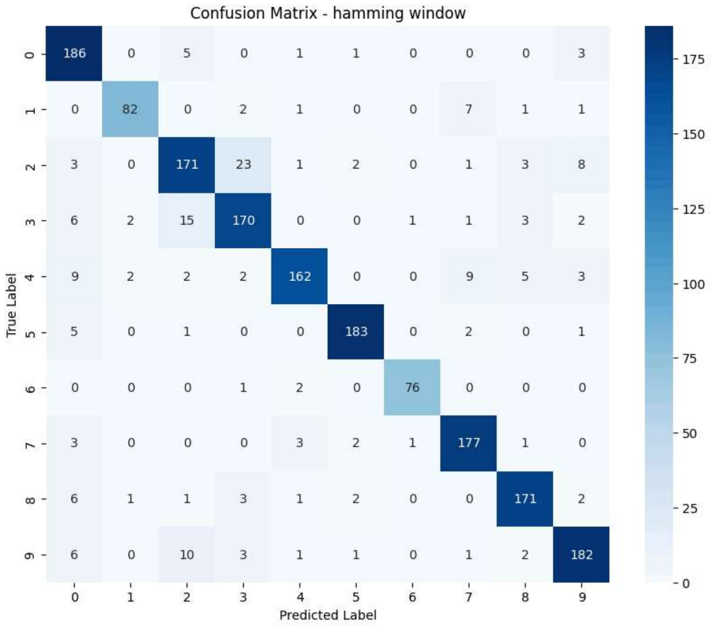
After training, model performance is assessed using:

- **Accuracy**: Percentage of correct predictions.
- **Precision & Recall**: Measure classification quality.
- **F1 Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Displays misclassified examples.

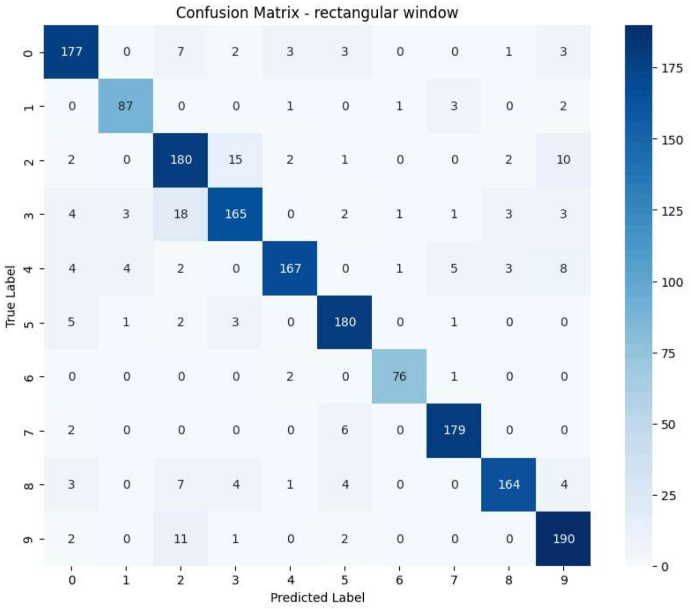
Confusion Matrix of different windowing method



	0	1	2	3	4	5	6	7	8	9
0	184	0	3	0	1	2	0	1	0	5
1	0	86	0	1	2	0	0	3	0	2
2	7	0	173	22	3	2	0	0	1	4
3	3	2	10	174	1	1	1	1	4	3
4	1	1	0	1	181	0	0	3	2	5
5	5	0	1	1	0	181	0	0	3	1
6	0	0	0	0	3	0	76	0	0	0
7	5	0	0	0	3	4	1	174	0	0
8	4	1	5	2	0	3	0	1	171	0
9	5	0	9	2	2	1	0	0	2	185



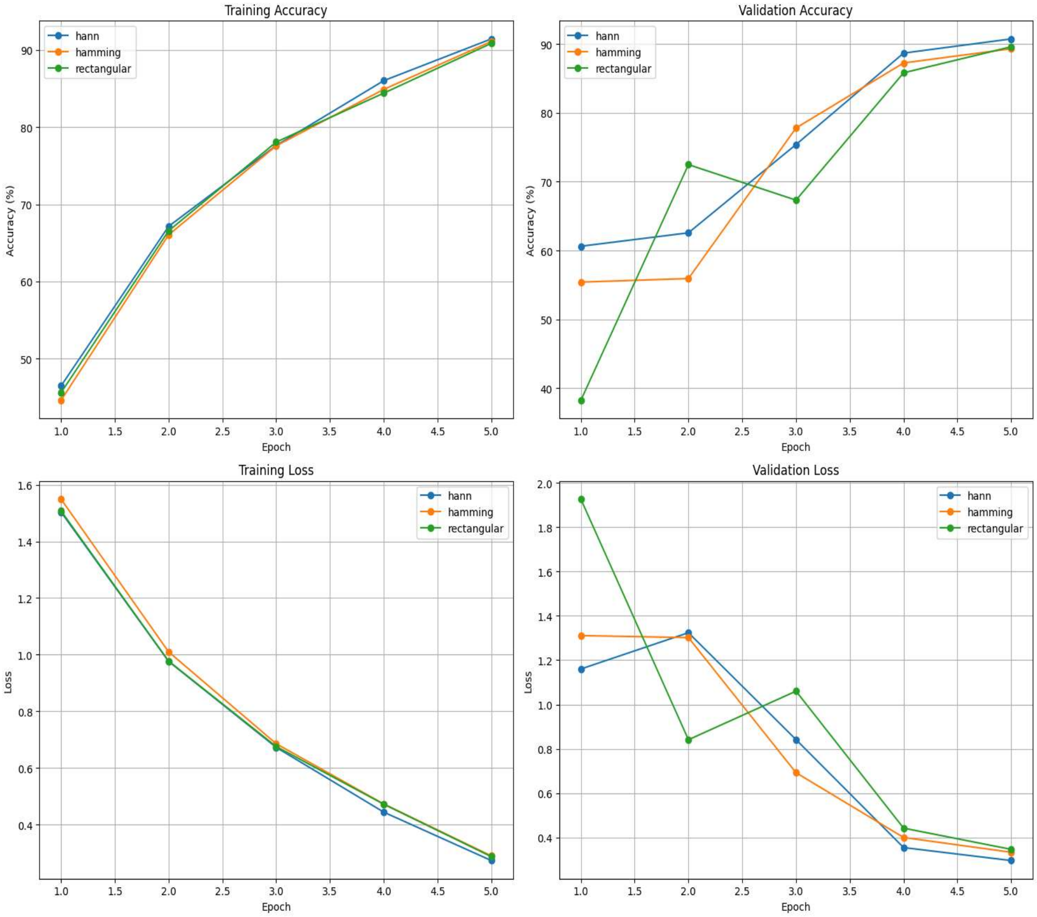
	0	1	2	3	4	5	6	7	8	9
0	186	0	5	0	1	1	0	0	0	3
1	0	82	0	2	1	0	0	7	1	1
2	3	0	171	23	1	2	0	1	3	8
3	6	2	15	170	0	0	1	1	3	2
4	9	2	2	2	162	0	0	9	5	3
5	5	0	1	0	0	183	0	2	0	1
6	0	0	0	1	2	0	76	0	0	0
7	3	0	0	0	3	2	1	177	1	0
8	6	1	1	3	1	2	0	0	171	2
9	6	0	10	3	1	1	0	1	2	182



	0	1	2	3	4	5	6	7	8	9
0	177	0	7	2	3	3	0	0	1	3
1	0	87	0	0	1	0	1	3	0	2
2	2	0	180	15	2	1	0	0	2	10
3	4	3	18	165	0	2	1	1	3	3
4	4	4	2	0	167	0	1	5	3	8
5	5	1	2	3	0	180	0	1	0	0
6	0	0	0	0	2	0	76	1	0	0
7	2	0	0	0	0	6	0	179	0	0
8	3	0	7	4	1	4	0	0	164	4
9	2	0	11	1	0	2	0	0	0	190

\

Here we are comparing the accuracy and training loss of all three models



The **UrbanSound8K** dataset was used to train a **ResNet50-based CNN model** for environmental sound classification. The models were trained using **three different windowing techniques (Hann, Hamming, Rectangular)**, each influencing the **spectrogram generation** differently. The results of training for **5 epochs** are compared below.

windowing techniques	**Val Accuracy**	**Train Accuracy**	**Precision**	**Recall**	**F1 Score**
Hann Window	0.9073	0.9144	0.9078	0.9073	0.9072
Hamming Window	0.8930	0.9111	0.8948	0.8930	0.8929
Rectangular Window	0.8958	0.9087	0.8976	0.8958	0.8960

The **Hann window** provided the **best overall performance**, with the **highest validation accuracy (90.73%)** and **best generalization**. The **Hamming** and **Rectangular windows** performed slightly worse, though still achieving competitive accuracy.

2. Windowing Technique Comparison

A. Hann Window (Best Performance)

- **Highest Validation Accuracy (90.73%)**: This indicates the model generalizes well to unseen data.
- **Training Accuracy (91.44%)** is **very close to validation accuracy**, showing **minimal overfitting**.
- **Balanced Precision, Recall, and F1 Score (~90.7%)**:
  - **Precision (90.78%)**: The model makes accurate predictions.
  - **Recall (90.73%)**: It correctly identifies relevant sounds.
  - **F1 Score (90.72%)**: The harmonic mean of precision and recall is the highest.

Why is Hann the Best?

- The Hann window minimizes spectral leakage, resulting in cleaner spectrograms with sharp frequency resolution.
- This helps the ResNet50 model extract meaningful frequency patterns, improving classification accuracy.

B. Hamming Window (Slightly Lower Performance)

- **Validation Accuracy (89.30%)**, **Training Accuracy (91.11%)**:
  - Shows **slight overfitting** (gap of ~2%).
- **Precision (89.48%)**, **Recall (89.30%)**, **F1 Score (89.29%)**:
  - All metrics are **slightly lower than Hann**.

Why is Hamming Worse than Hann?

- The Hamming window also reduces spectral leakage, but it retains more side lobes than Hann.
- This results in slightly blurrier frequency components in the spectrograms, making classification less precise.

C. Rectangular Window (Intermediate Performance)

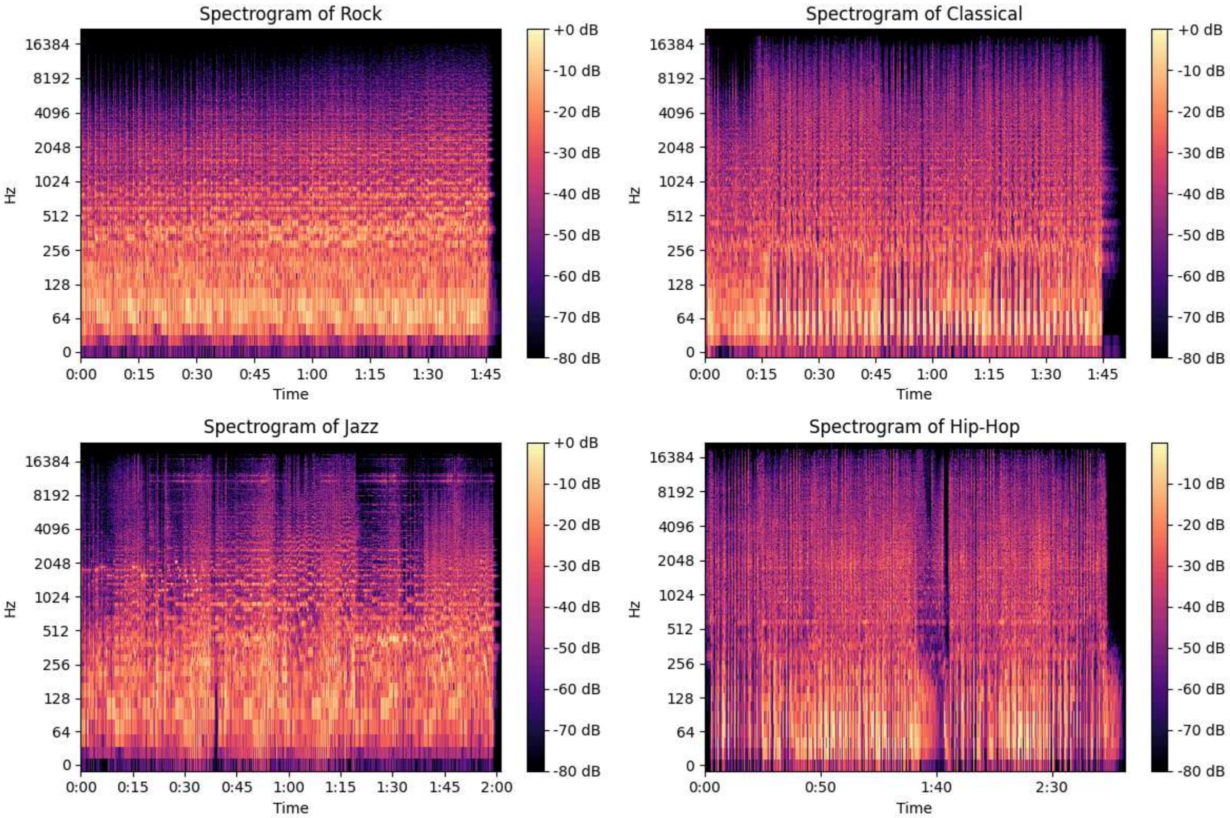
- **Validation Accuracy (89.58%)**, **Training Accuracy (90.87%)**:
  - Performs **better than Hamming but slightly worse than Hann**.
- **Precision (89.76%)**, **Recall (89.58%)**, **F1 Score (89.60%)**:
  - Better than Hamming but below Hann.

Why is Rectangular Worse?

- No smoothing effect → Leads to higher spectral leakage.
- This causes frequency smearing in the spectrograms, making it harder for ResNet50 to distinguish sounds accurately.

Task 2:

The provided spectrograms for four genres (Rock, Classical, Jazz, and Hip-Hop) give a detailed visual representation of their frequency content over time. Below is a genre-wise analysis:



1. Rock

- **Observations:**
  - The spectrogram shows significant activity across mid to high-frequency ranges (512 Hz to 8192 Hz).
  - There is consistent energy throughout the time duration, indicating a balanced mix of instruments like electric guitars, bass, and drums.
  - The intensity levels suggest a dense texture, typical of rock music.

- **Interpretation:**
  - Rock music often features distorted guitar riffs and prominent drum beats, reflected in the high energy across a wide range of frequencies.
  - The spectrogram's brightness in the upper-mid frequencies aligns with the aggressive and vibrant sound characteristic of rock.

2. Classical

- **Observations:**
  - Energy is concentrated in the lower frequencies (64 Hz to 1024 Hz), with smooth transitions over time.
  - Sparse activity in the higher frequencies, resulting in a darker spectrogram.
  - The dynamic range is wide, with quieter sections interspersed with louder ones.
- **Interpretation:**
  - Classical music often emphasizes string instruments, woodwinds, and pianos, which generate rich harmonic content in the lower and mid-frequencies.
  - The less dense high-frequency content reflects the absence of percussive and electronic elements.

3. Jazz

- **Observations:**
  - The spectrogram shows intermittent bursts of energy across a broad frequency range (128 Hz to 8192 Hz).
  - There is notable variation in intensity over time, indicative of improvisation and solos.
  - The distribution of energy in both low and high frequencies highlights the use of bass instruments and brass or woodwind solos.
- **Interpretation:**
  - Jazz's dynamic and rhythmic complexity is apparent in the variable energy patterns.
  - The presence of high frequencies corresponds to brass instruments, while the basslines dominate the lower frequencies.

4. Hip-Hop

- **Observations:**
  - Strong low-frequency content (64 Hz to 256 Hz), indicative of heavy basslines and beats.
  - Periodic repetition of patterns is visible, suggesting the structured nature of beats in Hip-Hop.
  - Sparse mid-to-high frequency activity compared to Rock and Jazz.
- **Interpretation:**
  - The emphasis on low frequencies reflects the genre's reliance on bass-heavy beats and sub-bass.

    - The rhythmic and repetitive nature is key to Hip-Hop, as seen in the spectrogram's temporal periodicity.

Overall Comparison

- **Frequency Range**:
    - Rock and Jazz display a broad frequency range, with significant activity in both low and high frequencies.
    - Classical and Hip-Hop focus more on specific frequency ranges—lower frequencies in Classical and bass-dominant frequencies in Hip-Hop.
- **Texture and Density**:
    - Rock and Jazz have a denser spectrogram, signifying complex instrumentation and layering.
    - Classical is more sparse but smooth, while Hip-Hop shows a rhythmically repetitive structure.
- **Dynamic Range**:
    - Classical exhibits the widest dynamic range, with visible quiet and loud sections.
    - Rock and Jazz maintain consistent energy, while Hip-Hop focuses on repetitive low-end patterns.

This analysis reveals how spectrograms can effectively capture the distinguishing features of musical genres. These visual differences align with the unique acoustic properties and instrumentation of each genre, providing insights into their composition and auditory characteristics.

References:

1. Task2_audio_link
2. GitHub
3. Wav2Small: Distilling Wav2Vec2 to 72K parameters for Low-Resource Speech emotion recognition
4. https://huggingface.co/dkounadis/wav2small
5. https://github.com/smitkiri/urban-sound-classification
6. https://medium.com/@aakash__/classifying-audio-using-pytorch-84861f3505ea
7. https://librosa.org/doc/latest/index.html
8. https://pytorch.org/