Spectrogram Creation, Windowing Techniques, and Classification Performance

Using the UrbanSound8k Dataset

1 Introduction

This report explores an audio classification experiment using the **UrbanSound8k** dataset. The study examines:

1. Generating spectrograms using the Short-Time Fourier Transform (STFT) with three different windowing techniques (**Hann**, **Hamming**, and **Rectangular**).
2. Training two classifiers (**SVM** and **MLP**) using features extracted from these spectrograms.
3. Comparing classification performance across different windowing methods.
4. Demonstrating that a **CNN** can achieve higher accuracy by leveraging spectrogram representations more effectively.

2 Dataset

The **UrbanSound8k** dataset is commonly used for classifying urban soundscapes and contains 10 classes:

- Air Conditioner
- Car Horn
- Children Playing
- Dog Bark
- Drilling
- Engine Idling
- Gun Shot
- Jackhammer
- Siren
- Street Music

The dataset is divided into 10 folds. A standard experimental setup involves training on 9 folds and testing on 1 fold.

3 Windowing Techniques

In STFT, an audio signal is divided into small frames, each multiplied by a **window function** before applying the Fourier Transform. Windowing helps **reduce spectral leakage** and shapes the frequency response. We tested three window functions:

3.1 Rectangular Window

- Also called the "boxcar" window.
- Applies uniform weight to all samples in the frame.
- Easiest to implement, but tends to have high spectral leakage.

3.2 Hamming Window

- A **tapered** window that reduces side-lobe amplitudes while preserving reasonable main-lobe width.
- **Spectrogram appearance:** Smoother than Rectangular window, with clearer harmonic detail.

3.3 Hann Window

- Another **tapered** window, closely related to Hamming but with slightly different taper shape.
- **Spectrogram appearance:** Often the cleanest in terms of harmonic trajectories and minimal spectral leakage.

3.4 Visual Comparison of Spectrograms

Spectrograms of a sample siren sound show:

- **Hann and Hamming Windows:** These produce smoother contours with less spectral leakage compared to the Rectangular window.
- **Rectangular Window:** This window introduces vertical stripes, reducing clarity in the time-frequency domain.
- **Hann Window:** Among these, the Hann window provides the cleanest harmonic structure, which can improve classification performance.

4 Methodology

4.1 Spectrogram Generation

1. **Load Audio:** Each clip is sampled (e.g., at 16 kHz).
2. **STFT:** Window size = 1024 samples; Hop size = 512 samples; Window function = Hann/Hamming/Rectangular.
3. **dB-scale Conversion:** Magnitude or power spectra are converted to decibel scale to form 2D time-frequency representations.

4.2 Feature Extraction and Classification

Spectrograms can be:

- Flattened directly into feature vectors, or
- Transformed via hand-crafted features like Mel-frequency cepstral coefficients (MFCCs).

Classifiers:

1. **Support Vector Machine (SVM):**
   - RBF kernel, $C = 1.0$, $\gamma =$ scale
2. **Multilayer Perceptron (MLP):**
   - 1 hidden layer (100 neurons), learning_rate = 0.001, max_iter = 1000

Additionally, a **CNN** was explored (*AudioCNN*), learning directly from spectrogram inputs.

5 Experimental Results

5.1 Window-Based Comparison (SVM, MLP)

Rectangular Window:

- MLP Accuracy: 60.81%
- SVM Accuracy: 64.28%

Hann Window:

- MLP Accuracy: 60.81%
- SVM Accuracy: 69.89%

Hamming Window:

- MLP Accuracy: 62.84%
- SVM Accuracy: 67.74%

5.2 Observations

- Hann and Hamming windows outperform Rectangular due to reduced leakage.
- SVM sees larger gains from improved windowing, reaching nearly 70% with Hann.
- MLP stays in the 60%–63% range, less sensitive to window choice.

6 CNN Performance

A CNN was tested with:

- Two convolutional layers (64 $\rightarrow$ 128 filters) + MaxPooling.
- A fully connected layer (1024 neurons) + Output layer for 10 classes.

Results:

- CNN on spectrograms: $\sim$ 80% test accuracy (10 fold) and $\sim$ 85% train accuracy (1-9 fold).

CNNs can learn local time-frequency features effectively, surpassing SVM/MLP.

7 Conclusions

- **Windowing Choice Matters:** Hann window yields best results ($\sim$ 70% via SVM).
- **SVM vs. MLP:** SVM capitalizes better on cleaner spectrograms; MLP remains more stable but at lower accuracy.
- **CNN Performance:** Deep learning on spectrograms achieves higher overall accuracy ($\sim$ 80%).

8 Future Work

- Systematic hyperparameter optimization (e.g., grid search) for SVM and MLP.
- Data augmentation strategies (pitch/time shifts, noise addition).
- Exploring deeper CNNs or Transformer-based models for further gains.