AXS230094 – Anand Singh

Project Report: Audio Classification Using MFSC and MFCC with Clustering

Objective: This project focused on distinguishing environmental audio sounds using Mel-frequency spectral coefficients (MFSC) and Mel-frequency cepstral coefficients (MFCC) as primary features. Clustering techniques were applied to categorize different types of audio, and the project explored how feature extraction parameters affected clustering performance, aiming to optimize settings for accurate classification.

Data Overview: The dataset included audio files representing diverse environmental sounds, such as crowd noise, motor sounds, and water flow. These categories allowed for examining sound features in a way that facilitated distinguishing between different types of sounds.

Tools and Libraries: The project utilized several Python libraries:

- Librosa for audio processing and feature extraction
- Scikit-learn for clustering and evaluation
- Matplotlib for visualizing results

Methodology

- Feature Extraction:
    - MFSC and MFCC features were extracted from each audio file to represent spectral and cepstral characteristics. These features were suitable for capturing frequency-based variations critical to classifying audio data. The impact of frame size, overlap, and number of Mel filters on feature quality was examined.

- Synthetic Time Series Generation:
    - Randomized segments of different class samples were used to create synthetic time series, allowing longer sequences with varied transitions between classes and improved representation.

- Clustering Algorithm:
    - An ART2 clustering algorithm was applied to categorize data points based on MFSC and MFCC features. Key parameters such as the vigilance threshold and buffer size played significant roles in clustering sensitivity. Label smoothing was introduced to stabilize classifications, minimizing short-lived misclassifications.

Challenges Faced

- Limited Data for Feature Extraction:
    - High-quality MFSC and MFCC feature extraction requires extensive labeled audio data, and limited data can affect the quality of extracted features, potentially resulting in underfitting.
- Solution:
    - To address this, synthetic data was generated by combining segments from each class, which improved data representation and feature diversity.
- Parameter Sensitivity:
    - Parameters such as frame size, hop length, Mel filters, and the ART2 vigilance threshold significantly impacted clustering outcomes. Each required tuning to optimize feature quality and clustering effectiveness.
- Solution:
    - Systematic parameter testing was conducted to identify optimal configurations, balancing feature detail and computational efficiency.
- Computational Complexity:
    - Larger frame sizes and reduced hop lengths increased the number of frames generated, leading to higher computational demands. This added complexity affected processing speed.
- Solution:
    - A compromise was reached by selecting frame sizes and hop lengths that maintained classification quality while minimizing computational demands.
- Noise and Label Smoothing:
    - Initial clustering results showed high noise levels, with frequent misclassifications at segment boundaries. Label smoothing improved classification stability but introduced a risk of over-smoothing.
- Impact of Different Parameters: Frame Size and Hop Length:
    - Larger frame sizes captured more spectral details but reduced time resolution, which could obscure transient features. Smaller hop lengths smoothed transitions in time series but increased computational requirements due to the higher frame count. The selected frame size of 30 ms and hop length of 10 ms offered a good balance between detail and efficiency.
- Number of Mel Filters (n_mels):
    - A higher number of Mel filters provided more detailed MFSC features but could make clustering more challenging for classes with

overlapping spectra. Reducing the number of Mel filters simplified distinctions but risked losing important details.

- ART2 Clustering Vigilance Threshold:
    - A higher vigilance threshold produced more clusters with finer separations, beneficial for capturing subtle spectral differences. Lower thresholds led to fewer clusters, simplifying the process but resulting in broader classifications. A threshold of 0.7 was found to work well with this dataset.
- Label Smoothing Window:
    - A smaller smoothing window effectively reduced noise without overly blurring transitions between classes. Larger windows excessively smoothed labels, impacting classification precision during transitions.

Results and Analysis: Accuracy and Confusion Matrices:

The clustering results were assessed using accuracy scores and confusion matrices for both MFSC and MFCC features. MFSC showed higher classification accuracy for classes with unique frequency characteristics, while MFCC performed better for complex sounds.

Visualizations: Gantt-style plots were used to show cluster assignments over time, revealing the effectiveness of MFSC and MFCC features in identifying transitions between sound classes. These visualizations offered insights into clustering behaviour.

Optimal Parameter Configuration: The configuration of a 30 ms frame size, 10 ms hop length, 40 Mel filters, and a vigilance threshold of 0.7 provided an effective balance between clustering accuracy and computational efficiency.

Conclusion: The project demonstrated the utility of MFSC and MFCC features for audio classification. The accuracy and stability of clustering were influenced by frame size, hop length, Mel filters, and vigilance threshold settings. Parameter testing and adjustments significantly improved the final clustering outcomes.