MLPC Report

Team IMPORTED

Mahmut Öztürk **Michele Giambelli** **Barış Bakırdöven** **Korhan Erdoğdu**

Contributions

Mahmut Öztürk checked data consistency and quality by selectively listening to sample audio and analyzing via python libraries. Michele Giambelli worked with label characteristics to identify class unbalances. Barış Bakırdöven analyzed feature characteristics to determine relevant/redundant features. Korhan Erdoğdu looked at feature/label agreements to see which features would be most useful for classification. Everyone contributed to writing report and presentation.

1 Data Consistency and Quality

1.1 For point A;

```python
import speech_recognition as sr
import os

# Adjust the folder path to where your .wav files are located
klasor_yolu = 'MLPC24_speech_commands_raw_waveforms/Brötchen'

# List all .wav files in the specified folder and limit to the first 200
wav_dosyalari = [f for f in os.listdir(klasor_yolu) if f.endswith('.wav')][:200]

# Create a speech recognition object
r = sr.Recognizer()

# Initialize containers for outcomes
recognized_words = []
unrecognized_files = []
error_files = []

# Loop through each .wav file and attempt to convert speech to text
for dosya in wav_dosyalari:
    dosya_yolu = os.path.join(klasor_yolu, dosya)  # Full path to the file
    try:
        # Load the audio file
        with sr.AudioFile(dosya_yolu) as kaynak:
            audio = r.record(kaynak)  # Read the entire file

        # Use Google's Web Speech API to convert the audio to text in German
        metin = r.recognize_google(audio, language='de-DE')
        print(f"{dosya}: {metin}")
        recognized_words.append(metin)
    except sr.UnknownValueError:
        print(f"{dosya}: Speech could not be recognized.")
        unrecognized_files.append(dosya)
    except sr.RequestError as e:
        print(f"{dosya}: Could not request results from the service; {e}")
        error_files.append(dosya)
```

Figure 1: Speech Recognition Code

In Figure 1 the Python code utilizes the speechrecognition library to convert speech from multiple .wav files, particularly useful for batch processing multiple audio files and converting speech to text. It handles cases where speech cannot be recognized or errors occur during processing. This way unrecognized speeches are found faster and checked manually by the members as well.

There are some inconsistencies within the dataset, notably instances where recordings are prematurely cut off before the completion of words. For instance, in the second example of the 'Staubsauger' audio file, the recording terminates before the word is fully articulated. Additionally, there are occurrences where audio recordings commence after the

initiation of speech, resulting in truncated utterances. Furthermore, within the dataset, there are instances of recordings that lack any verbal content yet have been categorized within specific word folders. An exemple of this is the audio files numbered 1170 and 979 within the 'alarm' folder, wherein no discernible speech is present despite its placement within the designated category. We have used SpeechRecognation library to recognise the "Speech could not be recognized." labeled data fastly and also checked them manually by listening them.

1.2 For point B;

After listening to some of the recordings, there are some biases in the data set. As exchange students, our awareness of dialectal differences was somewhat limited, but upon consulting with local individuals, we were informed of noticeable variations across German dialects, presenting a potential bias within the dataset. Furthermore, the inclusion of foreign student voices within the dataset could also introduce bias. Upon initial examination of a subset of the data, it becomes apparent that the majority of the audio recordings feature male students which plays a role to make a bias in the data set as well.

1.3 For point C;

Typically, these audio records serve to capture the commonplace sounds found within a household environment. Among the auditory elements commonly encountered in such recordings are instances of rustling and crackling, the friction of fabric, the sound of objects hitting a table, the distinct clicks of keyboard keys and mouse buttons, taps echoing from the kitchen or bathroom, occurrences of items being dropped, the occasional whistle, the steady rhythm of water dripping, the sound of someone drinking water, moments of yawning, instances of heavy breathing, background music, incoming phone calls, and occasional coughing fits.

2 Label Characteristics

The 20 words in the dataset are divided in 10 words that are useful for our classification project. Eight of them are referred to devices: Fernseher (TV), Heizung (heating), Licht (lights), Lüftung (ventilation), Ofen (oven), Alarm (alarm), Radio (radio), Staubsauger (vacuum cleaner). Instead, 2 are commands for turning on and off the devices: an (on) and aus (off). Than, the other 10 are words that are not useful for the machine learning algorithm. Some of them are similar in pronunciation to the first 10 words. They are Brötchen (bread roll), Spiegel (mirror), wunderbar (wonderful), kann (can, similar to "an"), Haus (house, similar to "aus"), nicht (not, similar to "Licht"), warm (warm, similar to "Alarm"), offen (open, similar to "Ofen") etc. Also, in the dataset there are files that are labelled as 'other'. These means that the audio registration probably it is a noisy or speech.

2.1 For point A;

For the classification task it is useful to classify this different labels, into some classes. The idea is to split them into 4 different classes: device word, command, useless word and other. In the device word class we have the first eight words referred to devices. Then in command class the words aus and an. The class other is the same of the label. In the end, the class useless word contains all the ten words that are not useful. This could be useful for knowing how many audio registration contains useful words for the model and how many are only noisy. What we get is the following distribution of the snippet in the new four classes [fig. 2].

2.2 For point B;

The resulting four classes are very unbalanced between them. In particular, there is a slightly difference between the useless word class and device word. The first one contains 45.07 of the audio registration and the second one 36.06. However, there is a big difference between these first ones and the other two. Classes other and command together are only the 19.8 of the dataset. The words an and aus (class command) are present in the dataset only in 4982 audio (9.01).

3 Feature Characteristics

As part of the exploratory data analysis, feature characteristics must be investigated. By feature characteristics it is meant the relationship between features of the audio data. The dataset given contains 3-dimensional information, where rows are audio samples with a total number of 45296, columns are features which are total of 175, and the 3rd dimension is timesteps with 44 steps. The feature characteristics analysis focuses on the distribution of every single feature, with hopes that some highly correlated/redundant features can be found and then be dealt with in the data preparation step,

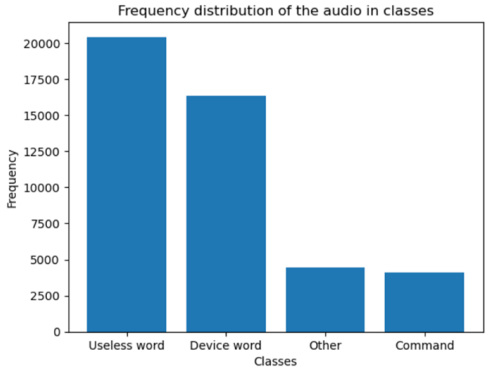

Figure 2: Frequency distribution in classes

which would reduce the computational load in the model training step. The features themselves are obtained from the audio data through different techniques such as Fast Fourier Transform. Even though application these techniques are not within the scope of this project, it is important to realize the existence of such a feature extraction from raw data step before the dataset was given. Within the 175, there are features related to bandwidth, contrast, flux...

3.1 For point A;

The first question to answer is how are these audio features distributed? The distribution of each feature is examined for all samples in all timesteps. This distribution is then visualized as histogram plots, where the x-axis is the numerical values that that feature takes, and the y-axis is frequently that value showed within all samples in all times. Distributions of first 12 features can be seen in the figure 3. The first impression here is that even though there are 175 distinct features, many of them belong to the same family, and those features look very similar to each other. This is the kind of strong correlation between features to look for and eliminate later. For example, contrast0,1,2,3,4,5,6 all looks almost the same (chi squared distribution) with minor differences when compared to difference between bandwidth0 and contrast0.

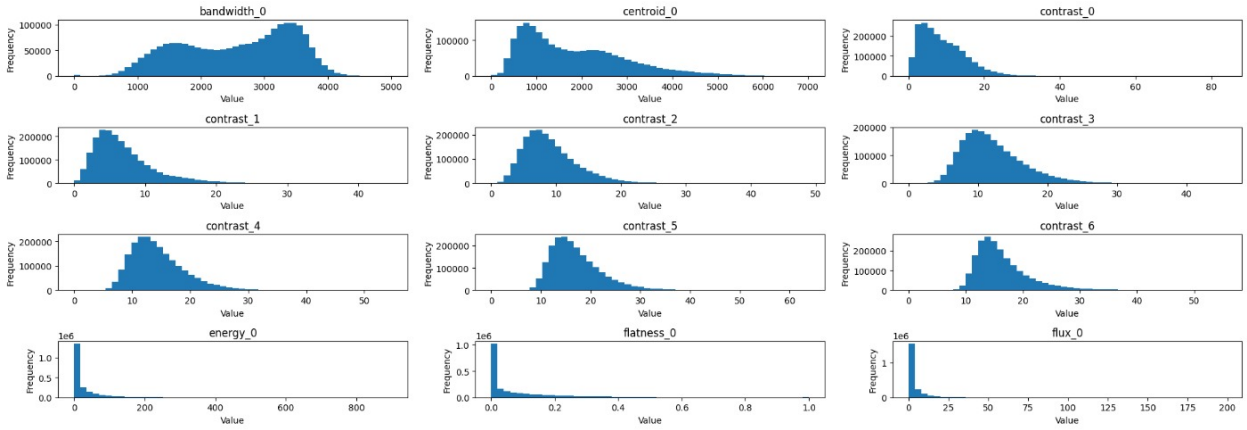

Figure 3: Distribution of first 12 features

3.2 For point B;

The second question at hand is related to the correlation between features. Even though the visual inspection of the feature distribution itself show strong correlation between features from same family, a correlation matrix is built and visualized for this task to determine if this applies for all families. This visualization for all features can be seen in figure 4, where the red indicates strong correlation and blue weak. The horizontal line of red dots is expected, since features are obviously strongly correlated with themselves, but the red square on top left indicate very strong correlation within the melspect feature family.

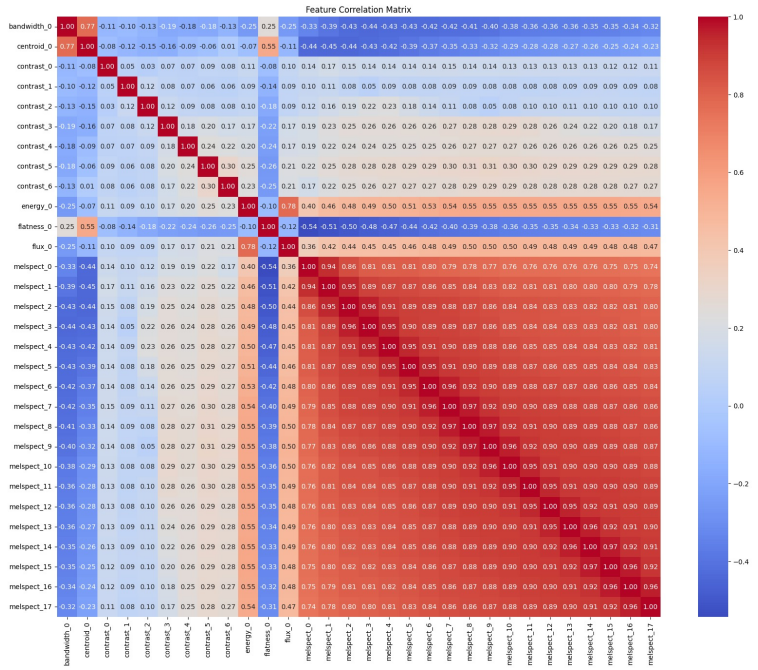

Figure 4: Correlation

3.3 For point C;

As the last part of this task, different speakers and their feature distributions are checked. 3 speakers are selected with id's 1, 50, and the visualization can be seen in figure 5 with the order of colors blue, orange, green. The distributions seem to have the same type (normal, chi-squared...), with slight differences in variance or mean that can be caused by different tones, accents, noise etc. The existence of such differences will beneficial for the model training.

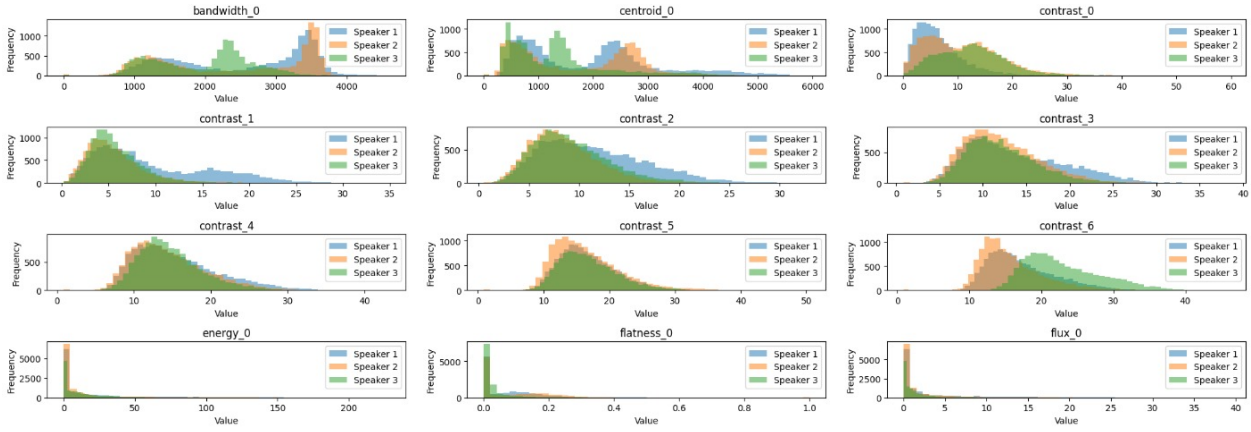

Figure 5: Feature distributions of 3 different speakers

4 Feature / Label Agreement

4.1 For point A;

Top 100 features with the highest variability have been investigated;

The features with the highest variability are mostly yin_{0_TXX} features at various timestamps. High variability in a feature across our dataset implies that the feature values differ significantly from one sample to another. This can be an

indication that the feature captures meaningful differences in the audio signals that could be useful for distinguishing between different words or sounds. Features with high variability might be particularly useful for classification tasks because they suggest the presence of distinct patterns or characteristics in the data that can differentiate between classes (in our case, possibly different words or sounds).

The fact that yin_{0_TXX} features dominate the top variability list suggests that pitch-related characteristics of the audio recordings vary widely across our dataset.

4.2 For point B;

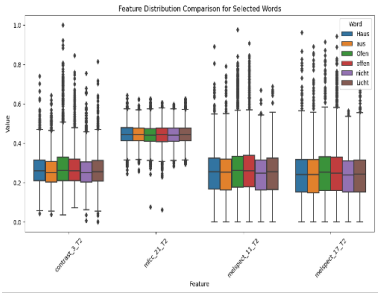

Figure 6: Frequency distribution in classes

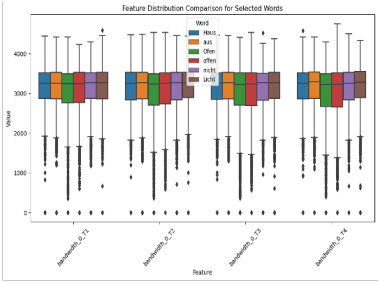

Figure 7: Frequency distribution in classes

The median value for each word is indicated by the line within each box. As an example, the feature distributions of "Haus - aus, Ofen - offen, or nicht - Licht" have been calculated for the timestamp two for the 4 features. The boxplot illustrates the fact that there is a high similarity between Haus and aus in the feature $melspect\_17\_T2$, however, we couldn't see that high similarity between these two in the feature $contrast\_3\_T2$. Therefore, it shouldn't be generalized that there will always be a similarity if both words contain similar letters. However, as a result, there does not seem to be a drastic difference in the distributions of the features.

For a futher understanding, same samples have been calculated but with the different feature and timestamps. It seems the median values for "Haus" and "aus" are somewhat similar across the timestamps, which suggest a degree of similarity in their feature distributions. However between the words that have no much letters in common, tend to differentiate in terms of their feature distributions, i.e aus and Ofen.

5 Conclusion

In conclusion, exploration of audio data for speech command recognition allowed Identification of inconsistencies and biases within the data, which implies the need for data augmentation. Further analysis revealed important feature correlations, indicating strong potential for dimensionality reduction of the data without losing prediction power by getting rid of redundant features. Key features demonstrating strong correlations with labels were highlighted, showing their importance for classification. Variations in feature distributions across speakers emphasized the need for diverse training data to enhance model robustness. These findings will be most important for the data preprocessing/preperation step and model design step, and our collaborative effort has laid a solid foundation for addressing the challenges of speech recognition in diverse conditions.