Vowel Recognition

October 8, 2022

1 Introduction

Speech recognition is a technology that has had a lot of improvement in recent years. While common knowledge indicates that consonants are the most dominant features in recognizing words, sometimes words consist of purely voiced utterances. For instance we can not distinguishing between words like "man" and "moon" by considering only the consonants. We can see that the proper speech recognition also requires robust recognition of voiced features. In this report, we explore methods to differentiate all the vowels from each other using machine learning methods. Our experiments focus on classifying unambiguously uttered vowels "a", "e", "i", "o" and "u".

The proper goal of this project is creating a decent vowels classifier. In Section 2, we formalize the vowel recognition problem into a machine learning problem, and furthermore in Section 3, we explain our methods of data refinement.

2 Problem Formulation

The commonly recognized features of a voiced utterance are intensity, pitch and timbre. Out of these features timbre, or the "fine detail", is the one that we can use to differentiate between vowels. When pressure is plotted as a function of time, timbre is defined by the shape of a single wavelength. On the frequency side timbre is encoded in the energy distribution. We will use the latter representation of sound waves as our features.

A Fourier transform of a sound wave expresses it as a sum of sine waves. Furthermore, the energy of a sine wave is proportional to the squares of the the amplitude and frequency. In summary, it is as follows:

$$E \propto A^2\omega^2 \tag{1}$$

One idea is to use the logarithmic energy distribution of the utterance's Fourier transforms between the frequencies 60 and 6000 Hz sampled at 60 Hz intervals as our sole feature. This one feature can be seen as a composition of 100 "smaller" features. Our labels will be the letters a, e, i, o and u. Since we are working with labeled data, our task falls to the category of supervised learning.

2.1 Dataset

We are using the DATASET_OF_VOWELS that can be found in Kaggle [1] as our base dataset. The base dataset consists of roughly 1600 labeled utterances from multiple different speakers. The utterances are vowels and are evenly distributed among a, e, i, o and u.

3 Methods

3.1 Dataset parsing

The base dataset requires heavy processing so that it fits our problem formulation. Here is a brief description of our processing methods: We split the original audio files to suitably sized time windows and take the Fourier transforms of those windows. We use a pitch detector to filter out unvoiced datapoints and pick ten samples from each audio file. The energy distributions of the samples are stored in csv format. For more details, see the bottom part of parser/tasks/parser.cpp in the Github repository [2]. We obtain Figure 1 when we plot the mean energy distributions of the datapoints. From the medians alone we can determine that our choice of features will be sufficient since the shapes are so different. In the feature space this corresponds to the clusters being far apart from each other.

In the end, we ended up generating two datasets. One contains the energy distributions and the other contains only the amplitude distributions. Both consist of around 16000 datapoints evenly distributed between the labels. The datasets are further split to training sets and testing sets: 20% of an original set goes to the testing set and the rest goes to the training set. Since our data is evenly balanced among the labels we do not need a large test set. At the same time, a bigger training set reduces the risk of overfitting.

3.2 Logistic regression

Since the clusters' centers of mass are far apart from each other, using logistic regression should produce decent results. Therefore logistic loss is an easy choice of loss function in the training phase. We will use 0/1 loss in the testing phase and visualize the results as a confusion matrix. Based on the visualization of the dataset in Figure 1, we expect the model to clearly recognize "a", "e" and "i" while "o" and "u" will get confused more often. This is indeed what happens as we can see from the confusion matrix in the Figure 2. The total accuracy of this model turns out to be 85% with the amplitude dataset. This is significantly higher than the result we obtain with the energy dataset, which scored 80% in accuracy.

References

[1] DATASET_OF_VOWELS, available at
https://www.kaggle.com/datasets/darubiano57/dataset-of-vowels?resource=download.

[2] Github repository, available at
https://github.com/UnilK/ML-project.

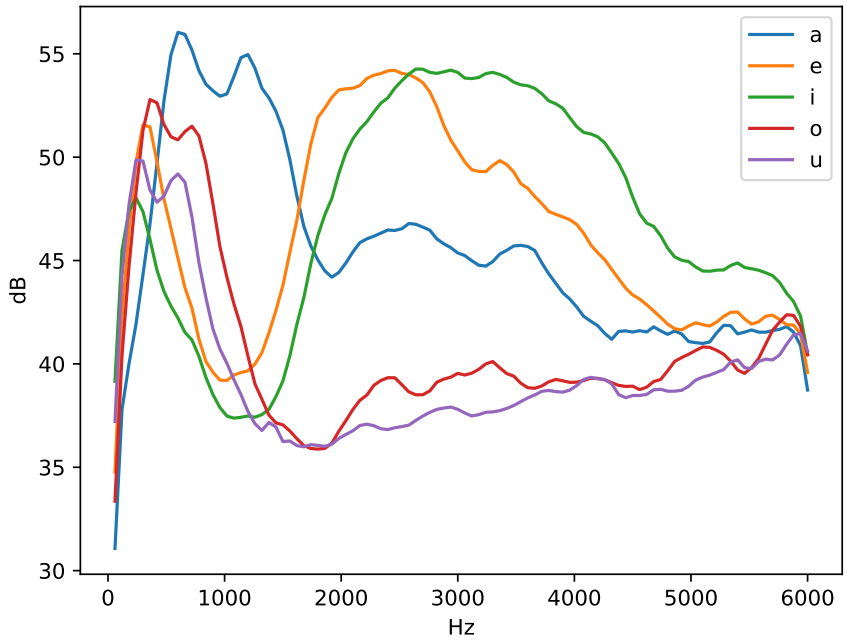

Figure 1: The mean energy distribution of vowels

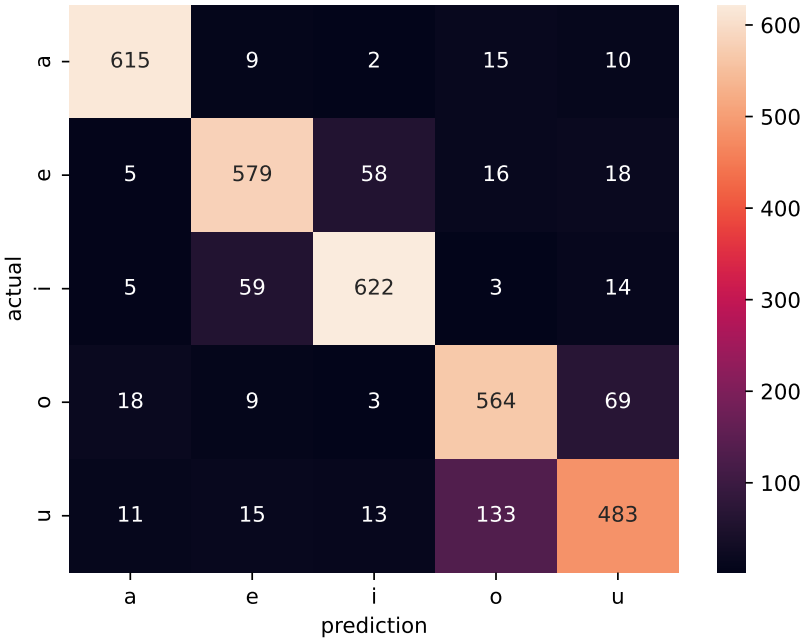

Figure 2: The confusion matrix obtained with logistic loss