Vowel Recognition

October 9, 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 distinguish between words like "man" and "moon" by considering only the consonants. We can see that proper speech recognition also requires robust recognition of voiced features. In this report, we explore methods to differentiate all 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 to create 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 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.

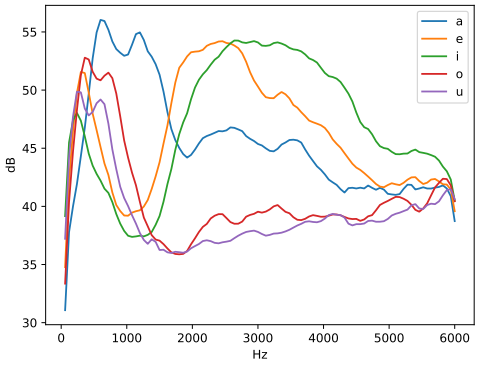


Figure 1: The mean energy distribution of vowels

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 into 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 into 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 for logistic regression and in all the other methods we will be using.

3.3 Decision Tree Classifier

Since the relation between the features and labels is non-linear, it would be essential to find a non-linear model. Due to implicit performing feature selection, requiring less effort for data preparation, eligibility to multiple feature values, and being capable of handling categorical data, leads us to use a decision tree classifier.

3.4 Neural Network

The main key in using a neural network is its flexibility which can perform well for both regression and classification problems, and it is outstanding to model with non-linear data with a large number of inputs like the situation we have here. Artificial neural networks have been proven to be universal approximators, so they should work well for our problem too. We can make our dataset fit the model with a very small change: we one-hot encode the labels. We can use squared error as our loss function as it is readily implemented in the python library that we are using. In the case that our hypothesis space is a constant value, the squared error produces the average of the datapoints. Interpolating this fact to other hypothesis spaces, we see that the squared error loss produces a result that is analogous to the best average of some kind. During the evaluation phase, we can encode the results back into letters by choosing the output with the maximum value. We use 5 hidden layers of size 20 with the reLU function acting as our non-linearity.

4 Results

As we are using 0/1 - loss in the evaluation phase, we can visualize the results as nice confusion matrices. Based on the visualization of the dataset in Figure 1, we expect the model to clearly recognize "a", "e" and "i" while it is likely that "o" and "u" will get confused more often.

4.1 Logistic Regression

We will consider logistic regression as our baseline method. 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. As expected, the letters "o" and "u" are confused more often. Also, surprisingly "i" and "e" seem to be confused rather often, as we can see from 2.

4.2 Decision Tree Classifier

One of the critical factors in using decision trees is the depth of the decision tree. It is known that high depth could cause overfitting in our model and make it more complex. Thus, the first step in using a decision tree is finding the best value for the depth of our decision tree to avoid underfitting and overfitting. Therefore, we used mean squared error to compute training and validation errors for decision tree regression on the amplitude dataset for the degrees of range from 1 to 20. We present these errors for degrees from 5 to 10 in Table 1. It is concluded the depth of 8 gives the minimum amount of validation error along with an acceptable amount of training error, and then, we can observe an increase in validation error indicating that the model starts to overfit. The total accuracy of this model turns out to be 84% with the amplitude dataset. By doing the same approach on the energy dataset, the total accuracy of 82% can be obtained which is less than the accuracy obtained from the amplitude dataset.

4.3 Neural Network

Neural networks are somewhat similar to logistic regression in our case in the sense that the output values will be correlated to the shapes found in the input data. Therefore we expect neural networks to produce similar results to logistic regression but better since they have bluntly more

Degree	5	6	7	8	9	10
Training Error	0.627362	0.538571	0.401090	0.351654	0.274587	0.218057
Validation Error	0.681900	0.630526	0.538232	0.534648	0.540621	0.546595

Table 1: Training and Validation Errors for decision tree regression for degrees from 5 to 10

predicting power. By making the hidden layers significantly smaller than the input layer we reduce the chance of overfitting. Based on Figure 2, this model confuses "o" and "u" significantly less often and achieves an 88% accuracy score with the amplitude dataset, and the test error obtained from this method is equal to 0.387096 which is remarkably less than the other methods.

4.4 Comparing methods

Comparing the total accuracy of these three methods indicates that on a given validation set, the neural network is performing better, and confuses vowel letters less as compared to logistic regression and decision tree classifier. Thus, in this problem, neural network could be a adequate method here. Moreover, the decision tree classifier has some downsides. For instance, a small change in the data could cause a large change in the structure and shows its instability, and it also needs higher time to train the model. In the end, comparing the total accuracy and the test error indicates that the neural network is the best ML method for this problem.

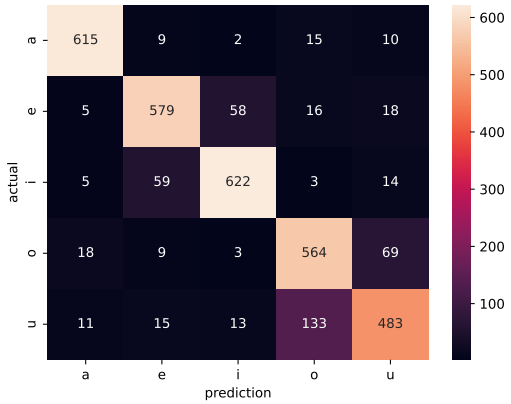
5 Conclusion

This report represents the application of machine learning in the problem of correctly distinguishing vowels and classifying them, which can have notable use in speech recognition technology. First of all, data processing is necessary, therefore, by taking Fourier transforms and preparing the energy distribution and the amplitude distribution of samples, we start to apply machine learning methods, logistic regression, decision classifier, and neural network. In applying these methods, 20% of the original set is considered as the test set and the rest is considered as the training set.

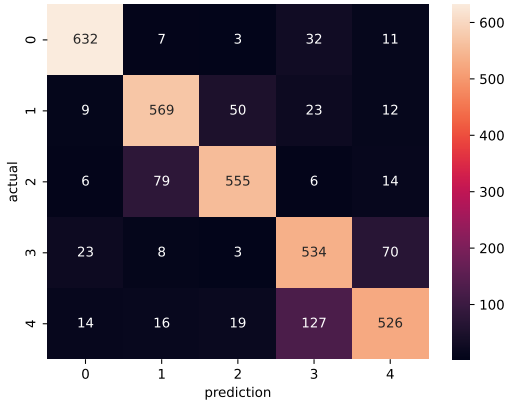
By using logistic regression, decision classifier, and neural network methods on the amplitude dataset, the total accuracy of 85%, 84%, and 88% can be obtained, respectively. Besides the high accuracy obtained from the neural network, once it is trained, the prediction is pretty fast.

Even though neural network gives us the promising result for this problem, there is plenty of scope for improvement. For instance, using other machine learning methods that are robust against noisy data, might be helpful and brings us notable results. Although neural networks usually perform as well as other methods of machine learning, it requires much more data than other traditional machine learning methods. Thus, the reliability of the neural network could be enhanced with a larger dataset.

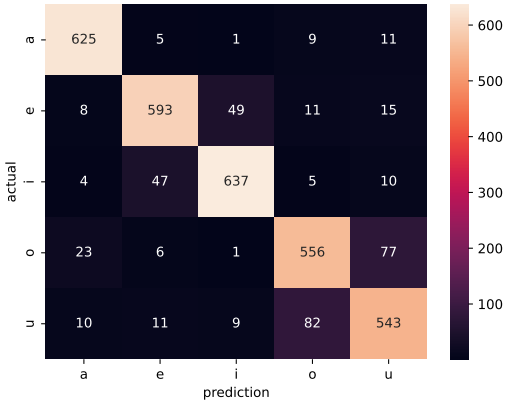
Furthermore, one of the research directions for future improvement could be considering additional features, like different accents which have effects on the amplitude distribution, we believe it would be undoubtedly beneficial to collect more data with a variety of recording environments, or even microphone positions.



(a) Logistic Loss



(b) Decision Tree Classifier



(c) Neural Network

Figure 2: The confusion matrices of the used ML methods.

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.