Intent Detection System Using Machine Learning Techniques

Mohammadreza Mashhadigholamali
Ali Samimi Fard
Politecnico di Torino

***Abstract*—In this report, we investigate possible methods to classify user intention. Our approach consists of using MFCC as a feature extractor and applying PCA as a feature reduction technique to the test dataset. Selected features go through Random Forest, Support Vector Machine, and Multilayer perceptron algorithms. ANN outperforms other methods with an accuracy equal to 88.9 percent**

I. PROBLEM OVERVIEW

Due to significant advances in the field of Natural Language Processing (NLP), voice command controls are becoming increasingly popular in daily life, for instance, virtual assistants, online helpdesks/chatbots, robot instruction, and so on. Obviously, employing an accurate intent classification technique is fundamental to delivering users a reliable and enjoyable experience. In this project, we attempt to detect and classify speakers' intentions using the "development" part of the dataset, including 9854 recorded voice files divided into seven categories.
Initially, some issues with this dataset must be appropriately addressed unless we get an unsatisfactory result at the final stage. First, most of these recorded voices are sampled at 16kHz, but we resample all files to have equal sampling frequency to eliminate the effects of this variable. Second, some recorded voices have unexpected silence at the beginning or end. Figure 1 shows the histogram of the sample files' duration. Accordingly, it is recommended to remove these extra parts. Then, we add some zero value to the end of the data which have lower length. In this case, we obtain recorded samples with equal time duration, allowing us to use classification models which need an equal number of features, such as Artificial Neural Network (ANN).
Additionally, In this dataset, the number of samples for each class is not equal, so it is necessary to tackle this problem to avoid imposing bias on the result. Therefore, we randomly oversample minority classes to provide balanced dataset. [1]

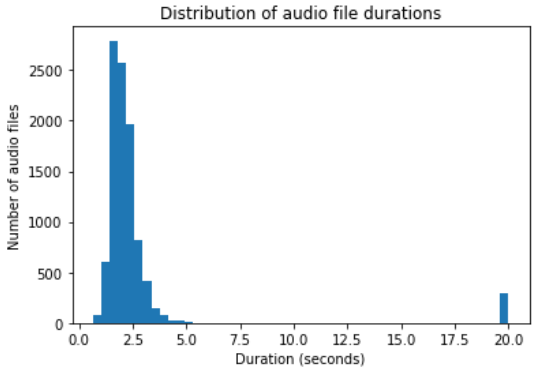

Fig. 1. Distribution of the durations of the recordings

II. PROPOSED APPROACH

A. DATA PREPROCESSING

In the preprocessing stage, we applied a noise reduction algorithm based on spectral gating to reduce time-domain noises. Figures 2 and 3 illustrate one random sample in a time domain before and after executing the padding and noise removal operation.

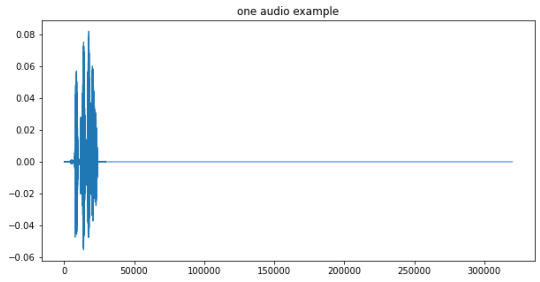

Fig. 2. Representation of a random recording in the time domain before preprocessing

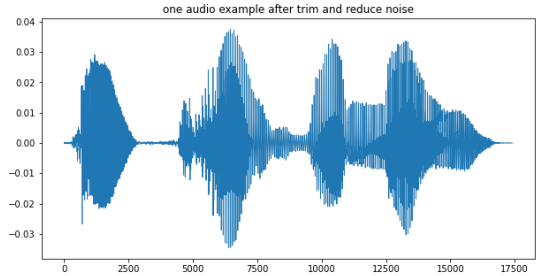

Fig. 3. Representation of a random recording in the time domain after preprocessing

We used Mel Frequency Cepstral Coefficients (MFCC) as a feature extractor for this assignment. MFCC is a renowned approach in the field of audio signal analysis due to its good performance, flexibility, and robustness to environment

artifacts. Also, Since speech is a non-stationary signal, short-term spectral analysis is the most common way to characterize the speech signal [2].
In this algorithm, frequency converts from Hz to Mel Scale. Since After executing the MFCC algorithm, we obtain a large number of features; it is necessary to use a feature selection technique to reduce the dimension feature vector. As a result, we used Principal Component Analysis (PCA).
PCA is used as a technique for data reduction/compression without any loss of data points. The purpose of PCA is to obtain a small number of principal components of a set of variables that retain information from the original data variables [3].
Finally, obtained values were standardized using the Min-Max algorithm, which scales all values in the range of [0, 1]. Figure 4 shows the block diagram of the proposed approach for intent detection.

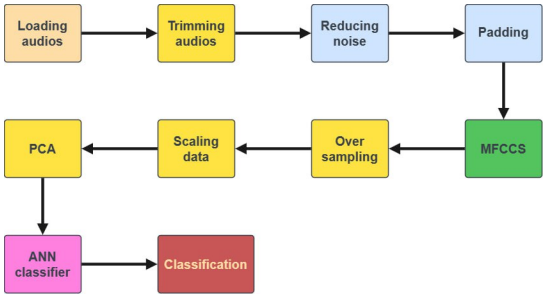

Fig. 4. Block diagram of the proposed approach for this project

B. MODEL SELECTION

Classification is the next step that comes after feature extraction. This step was completed using Machine Learning and Neural Networks algorithms in the following:

- Random Forest: Random forest is a supervised classification algorithm that is based on Decision Tree. This algorithm can produce excellent results, even without hyperparameters adjustment and has been widely used in the field of voice signal analysis [4].
- SVM: Support vector machine is also one of the most powerful classifiers in the field of voice recognition.
- ANN: Artificial neural network is a supervised machine learning algorithm applicable in various areas. This model has been studied in previous researches and has provided outstanding results [5].

C. HYPERPARAMETER TUNING

Some hyperparameters of each algorithm are chosen based on previous studies' results, consulting with domain experts, and trial and error. Other parameters did not modify manually. We build a 6-layer multilayer perceptron (MLP) with 1024, 512, 256, 128, and 64 hidden layers with the ReLu activation function. Also, since we have seven categories, choosing the number of output layer neurons equal to 7 is necessary. Softmax function was chosen as the activation function of the output layer.
Furthermore, we implemented a dropout layer between each layer to prevent overfitting in deep neural networks [6].
For SVM, it is crucial to choose an appropriate kernel function. We chose the Radial Basic Function (RBF) kernel, which performs a low to high dimensional feature transformation. This transformation allows non-linearly separable data to be linearly separable at a higher dimension [2].
Random State is the other parameter that affects the models' performance. If we set this parameter equal to *None* at the splitting data into test and train set step, we get different datasets across different executions. This results in getting dissimilar performances at the classification stage. In this assignment, we set this parameter to 0 [7].

III. RESULT

In this project, we classified audio data using three different algorithms. The experiments were done on a personal computer with an Intel Core i7 CPU and 16 Gigabytes of RAM. In order to evaluate the performance of the models, we used the "Evaluation" part of the dataset, including 1455 unlabeled samples. Table I shows the accuracy rate on the test dataset using RF, SVM, and ANN algorithms.

Model	Accuracy
Random Forest	33.9
SVM	79.4
ANN	88.9

TABLE I
THE HIGHEST OBTAINED ACCURACY IN THIS PROJECT WITH DIFFERENT CLASSIFIERS

It can be seen from the above table that the SVM algorithm gives an acceptable accuracy rate of 79.4%, compared to RF, which produces 33.9%. However, the best accuracy is obtained using the ANN algorithm equal to 88.9%.
Figures 5 and 6 show the loss and accuracy of the ANN classifier for test and train datasets at each epoch.
Finally, Figure 7 demonstrates the confusion matrix of the ANN algorithm.

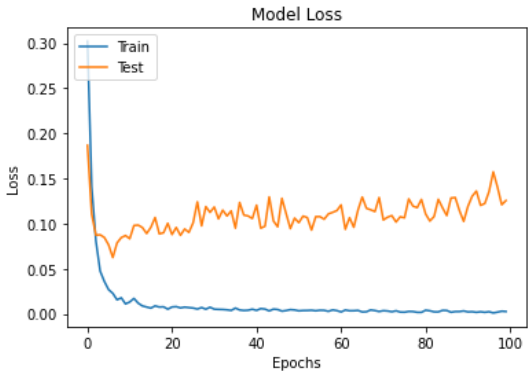

Fig. 5. ANN loss at each epoch

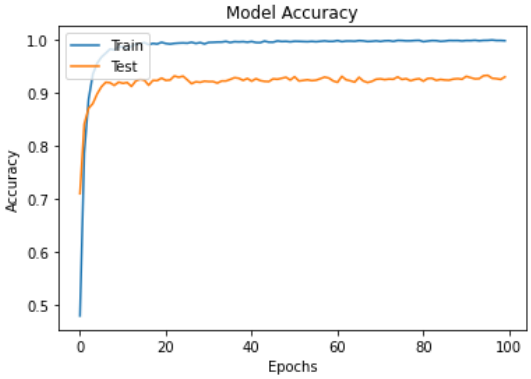

Fig. 6. ANN accuracy at each epoch

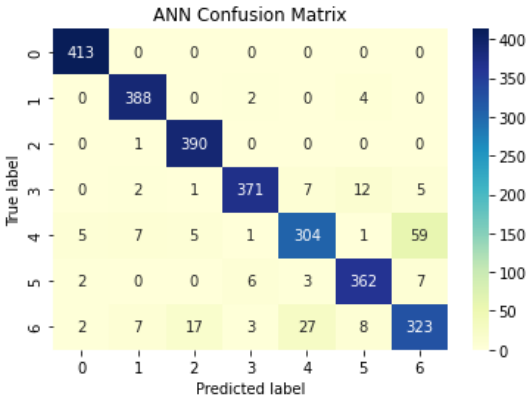

Fig. 7. ANN confusion matrix

IV. DISCUSION

In this assignment, we investigated three different algorithms for voice intent detection, using MFCC features and the PCA feature reduction algorithm. The best acquired accuracy was 88.9 percent with the MLP algorithm.
Furthermore, SVM yields an acceptable result compared with the ANN model, equal to 79.4 percent. In contrast, Random forest did a poor performance and achieved a 33.9 percent of accuracy. It can be concluded that this feature extraction algorithm should not be used along with this classifier.
In the following, there are possible approaches which worth investigating for future projects:

- Implementing other feature extraction methods like Mel spectrogram or Chroma features/chromagram
- Using other algorithms for classification, like convolutional neural network (CNN), K-nearest neighbor (KNN), and gradient boosting (GB)
- Implement optimization algorithms to find the best possible values for models' hyperparameters

REFERENCES

[1] (2021) Random oversampling and undersampling for imbalanced classification. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification.

[2] F. Abakarim and A. Abenaou, "Voice gender recognition using acoustic features, mfccs and svm," in *Computational Science and Its Applications–ICCSA 2022: 22nd International Conference, Malaga, Spain, July 4–7, 2022, Proceedings, Part I*. Springer, 2022, pp. 634–648.

[3] K. Jain, A. Chaturvedi, J. Dua, and R. K. Bhukya, "Investigation using mlp-svm-pca classifiers on speech emotion recognition," in *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, 2022, pp. 1–6.

[4] H. Jupalle, S. Kouser, A. B. Bhatia, N. Alam, R. R. Nadikattu, and P. Whig, "Automation of human behaviors and its prediction using machine learning," *Microsystem Technologies*, vol. 28, no. 8, pp. 1879–1887, 2022.

[5] R. Rendyansyah, A. P. Prasetyo, and S. Sembiring, "Voice command recognition for movement control of a 4-dof robot arm," *ELKHA: Jurnal Teknik Elektro*, vol. 14, no. 2, pp. 118–124.

[6] M. Zakariah, Y. Ajmi Alothaibi, Y. Guo, K. Tran-Trung, M. M. Elahi *et al.*, "An analytical study of speech pathology detection based on mfcc and deep neural networks," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.

[7] (2023) Sklearn model selection train test split. https://scikit-learn.org.