UCLA
Dept. of Electrical and Computer Engineering
ECE M214A: Digital Speech Processing
Winter 2023

Speaker Region Identification

Project Description

Oral presentations and code submission are scheduled for **Wednesday March 15th** during class time 10am - 12pm. For MSOL online students, it is scheduled for **Saturday, March 18th** at **12-3 pm** by Zoom.

I. Introduction

In this project, we are interested in implementing a speaker identification system that predicts the city of origin of the speaker of a given utterance. Acoustic features such as MFCCs, LPCs. may be used to perform the task.

II. Data

The Corpus of Regional African American Language (CORAAL)[1] contains speakers each belonging to one of five different US cities: 1) Rochester, NY (ROC), 2) Lower East Side, Manhattan, NY (LES), 3) Washington DC (DCB), 4) Princeville, NC (PRV), or 5) Valdosta, GA (VLD). We selected only the utterances in the corpus with length greater than 10 seconds. In addition, a few utterances have been corrupted by a 10dB babble noise masker.
Your model will also be tested on a blind test set. The blind test set will consist of a different set of speakers from the above cities.
All the wav files provided to you have been sampled at 44.1kHz.

III. Project Codebase

The enclosed project link provides access to the following folders with utterances: train_clean, test_clean, test_noisy. The following python notebook[2] contains functions that extract features from the utterances in the above folders, and calculates the accuracy through the use of an xgboost based model. In addition, wrapper.m can be used to extract features in MATLAB and then subsequently load the features into python.
Note: Do not modify the classification model

Details: The baseline system consists of the following: MFCC features (13) extracted from the utterances and used for training the xgboost model. More detailed instructions on the specific libraries used for feature extraction and the model are present within the python notebook

IV. Objectives

Your task is to derive a set of features and implement them to predict the city of origin of the speaker of every utterance. You will train on clean data and test with 1) clean data, and 2) noisy data.
Note: Please don't use the noisy data for training.

V. Evaluation Metrics

The accuracy calculated by the python script will be used for evaluation. Along with the Accuracy, report other classification metrics such as the Confusion matrix of your model. Results on both clean and noisy data, as well as performance of the blind test set will be considered for evaluation.

In addition to performance on the test sets, the explainability of the success of the model is also considered. This can be quantified by the impact of different features chosen as inputs to the model and the reasoning behind their effectiveness.

VI. Instructions

A. Setting up the project

a. Download project package from Bruinlearn.
b. Unzip the compressed file.
c. Upload the 'project_data' folder to your google account
d. Make a copy of the colab notebook[2]
e. Open the notebook and run all the cells

B. Run custom features from Python

a. In the accompanying python notebook, edit the extract_features function to calculate your custom features directly
b. Run the model training and inference steps directly.
c. Repeat for other trials.

C. Run custom features from MATLAB

a. Run all the cells in the baseline python script
b. Run the cells to extract the list of files in the train and test sets.
c. Open the accompanying wrapper.m file
d. Change the function call being called to extract features in wrapper.m to your custom feature name.
e. Use the provided python function to read in the generated csv file into python
f. Run the model training and inference steps directly.
g. Repeat for other trials.

VII. Baseline Results

The baseline script should take around 30 mins to run

Dataset	**Accuracy**
Test Clean	**79.64%**
Test Noisy	**58.21%**

VII. Oral Presentations

There will be oral presentations by the different teams describing their work. Presentations should be planned by the team as a group.

VIII. Report and Code

The report (one per group) should include:

- Introduction (what is the problem/why is it important)
- Background (literature survey)
- Project Description (features, algorithm, implementation, results, average run times, etc.)
- Summary and Discussion (also ideas for future work)
- References (cited throughout the report)
- Figures and flowcharts generally help clarify the text.

The report should be 4-pages long and have the same format as the INTERSPEECH conference.

The code should be turned in on the day of the presentation. Comments at the beginning of each function should describe what the function intends to do.

To evaluate the robustness of your system, we will use speech from a different set of unseen speakers to evaluate the system performance. You will run your scripts on the unknown data and submit the scores. The final report may be turned in on **Sunday, March 19th**

Useful References

1. CORAAL dataset:
   Kendall, Tyler and Charlie Farrington. 2021. The Corpus of Regional African American Language. Version 2021.07. Eugene, OR: The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal.

2. Google Colab Notebook for feature extraction and classification:
   https://colab.research.google.com/drive/1k0c0fiYtTYtlyeQTvEKZ4L_usNm1X5Tr

3. Dialect Density Estimation in African American English:
   A. Johnson, K. Everson, V. Ravi, A. Gladney, M. Ostendorf, and A. Alwan, "Automatic Dialect Density Estimation for African American English," in Interspeech 2022, 1283-1287, doi: 10.21437/Interspeech.2022-796

4. The difficulties of working with accented audio:
   A. Koenecke, A Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, “Racial disparities in automated speech recognition.” Proceedings of the National Academy of Sciences Apr 2020, 117 (14) 7684-7689; DOI: 10.1073/pnas.1915768117

5. Techniques for accent classification:
   R. Huang, J. H. L. Hansen and P. Angkititrakul, “Dialect/Accent Classification Using Unrestricted Audio,“ in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 2, pp. 453- 464, Feb. 2007, doi: 10.1109/TASL.2006.881695

6. Methods for speaker identification:
   D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

7. Useful toolkit for speech processing and feature extraction:
   http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html