Abstract

Speech processing is essential to many ML/AI applications including voice recognition, speech transcription, and virtual assistants. However, technology watchdogs have raised concerns that the underlying algorithms and language models perform poorly on sociolinguistic groups underrepresented in training data. Specialized models trained exclusively on recordings from specific dialects would presumably perform better – if those models could be correctly chosen from a speaker’s early utterances. This study aimed to classify the dialects represented in a corpus of voice recordings of UK speakers based on their power-spectral features (i.e. MFCCs) as well as phonetic features (i.e. formants & their ratios). Phonetic features were far more useful than MFCCs, accounting for ~80% of top-ten features across all four classes. Logistic regression (regularized) and XGBoost were the best-performing models, achieving accuracy benchmarks of 0.55 across all four classes examined (null: 0.25). Further work would focus on exploring other formant-based features, and developing models specific to single genders.

Introduction

Speech processing is an important domain of artificial intelligence essential to speech transcription, virtual assistant commands, and speaker identification. Several teams have attempted to determine speakers’ identity (*diarization*), intent, and even emotions based on features engineered from speech waveforms. However, researchers have shown that state-of-the-art commercial applications, such as in Apple’s Siri and Amazon’s Alexa, perform less well on the speech of Black Americans and other minority groups. Universal access to the fruits of technological progress should be a priority in the research and development of these systems.

This problem can alternatively be framed broadly as accent recognition. Speech models in widespread deployment today are typically trained on samples from homogeneous speaker groups, and thus perform poorly for speakers from different sociolinguistic backgrounds. Language models trained exclusively on minority groups would likely perform better. This strategy could work especially well on speakers of English as a foreign language (EFL), where accented speech is almost inevitable. During an automated customer service or touchtone interaction, for example, if an accent is identified early, the caller could be funneled into the specialized model so that subsequent commands are captured more accurately: one model for L2 speakers from Mexico, a different model for L2 speakers from China, etc. Since this could be deemed a form of profiling, it is essential that accent recognition modules all feed into the same language model downstream, ideally trained on large text corpora in which sociolinguistic differences in diction are largely smoothed out.

Methodology & Tools

The dataset analyzed was the [Open-source Multi-speaker Corpora of the English Accents in the British Isles](https://aclanthology.org/2020.lrec-1.804/), a collection of over 18,000 high-quality audio of English sentences recorded in Great Britain and Ireland by volunteers having different regional accents. The recording scripts were curated specifically for accent elicitation—including personal and location names within the regions in question—and provide high phoneme coverage and comparable frequencies between groups. The authors themselves state, "the dataset is intended for linguistic analysis as well as use for speech technologies." Six accent classes are provided, each consisting of recordings from 3-57 volunteers of both genders (Table 1). Due to the small number of recordings labeled Midlands (England) and Irish, and the lack of female speakers for Irish, only the first four classes were analyzed in the present study.

|  |  |  |
| --- | --- | --- |
| **Accent Group** | **Speakers (male/female)** | **Entries** |
| 1. Southern | 57 (28/29) | 8,492 |
| 2. Northern | 19 (5/14) | 2,847 |
| 3. Scottish | 17 (6/11) | 2,543 |
| 4. Welsh | 19 (8/11) | 2,849 |
| *5. Midlands* | *5 (2/3)* | *696* |
| *6. Irish* | *3 (0/3)* | *450* |

Map

Description automatically generated Figure 1: Dataset attributes

Two types of features were considered. Power-spectral features included MFCCs (mel-frequency cepstrum coefficients; n=16 coefficients, default window length, sampling rate, etc.) and their derivatives (MFCC-delta, MFCC-delta2), and were calculated using Librosa. Since each file yielded as many MFCCs as windows for each coefficient, some form of aggregation was necessary. Therefore, each coefficient was averaged over the recording time to capture the sound energy in that part of the audible spectrum.

Phonetic features were obtained using Parselmouth, the Python implementation of phonological software Praat: namely, the fundamental frequency of the speaker’s voice in each recording (F0), as well as means and medians of the first four formants (F1, F2, F3, F4). Briefly, F1 corresponds to vowel height (tongue position in laryngeal cavity), F2 to vowel backness (tongue position with respect to palate), and F3 to rhoticity (‘r’ coloration) and lip rounding. Formants were scaled by F0 to normalize natural pitch differences between speakers. Formant ratios (F2/F1, F3/F1, F3/F2) were also included as features, since vowel perception also depends on their relative position. Note that since phonemes are well covered in all classes according to the authors, the usage of mean formants is intended to capture vowel height, backness, etc. as properties of speech generally, rather than specific vowels or sounds spoken in different regions of the UK.

Since each speaker provided dozens or hundreds of recordings, data were split such that no speaker was repeated in the train, validation, or test datasets. Experiments done without splitting showed excellent performance on MFCC features alone, but since it dropped considerably after splitting, it is assumed that this only amounted to voice matching (speaker identification).

Results & Discussion

Logistic regression (C=0.01) and XGBoost (optimizedby GridSearchCV) were the best performing models, each achieving a mean weighted accuracy of 0.55 on the four-group classification task, and F1 scores of 0.25~0.70 and 0.15~0.67 respectively on single-group (OVR) tasks. (Precision and recall were weighted equally: the cost of misidentification is presumed low in this context.)

Table

Description automatically generated

Coefficients and importances of formant-based features far outperformed MFCC-based features. Saliently, several of the most important features mapped intuitively to phonetic characteristics of the different accent/dialect groups of study:

* **F3/F1 (median), F3 (mean)**: F3 is determined by vowel rhoticity (‘r’-coloration) and lip rounding. Lower-frequency F3 indicates both greater rhoticity and rounding, but since F1 and F2 drop in tandem for rounded vowels, the ratio better captures rhoticity, while F3 itself is a better indicator of rounding. Northern and Scottish accents were associated with greater rhoticity, while Southern and Welsh accents were associated with less. Meanwhile, Scottish and Southern accents were associated with greater rounding, while Northern and Welsh accents were associated with less.
* **F2/F1 (mean)**: Refers to vowel compactness; in a global sense across all phonemes, maps to narrower variation in oral and laryngeal cavity size during speech. Dialects with greater variation might be described as ‘broad’ accents in vernacular. Northern/Scottish and Southern/Welsh accents were associated with less-compact and more-compact vowel pronunciation, respectively.
* **F1 (median)**: Refers to vowel height, where ‘high’ vowels are pronounced with the back of the tongue closer to the roof of the mouth. Vowel pronunciation tended to be slightly higher in Southern and Welsh speakers, while distinctively lower in Northern and Scottish speakers.

A screenshot of a computer

Description automatically generated with medium confidence

Sociopolitical issues with dialect profiling and discrimination are inherent in voice recognition technology. This project is framed at ensuring equal access to convenience and state-of-the-art technology, supplied by selecting the most appropriate of a variety of pre-trained models.