Features of speech audio for deep learning accent recognition

Yuvika Singh[1 [0000-0002-5430-3272]], Anban Pillay [1[0000-0001-7160-6972]], Edgar Jembere [1[0000-0003-1776-1925]]

[1] University of KwaZulu-Natal, Durban, South Africa
yuvikasingh@yandex.com

Abstract
An accent is the distinctive way words are pronounced. Every speaker has an accent, which varies by gender, age, formality, social class, geographical region, and native language. Accents differ by voice quality, phoneme pronunciation, and prosody. Since it is difficult to extract these exact features, existing work used alternate features. These features were generally spectral features, which captured the frequency of speech. Such features included the Mel-Frequency Cepstral Coefficient (MFCC), Spectrogram, Chromagram, Spectral Centroid, and Spectral Roll-off, which were extracted from raw audio samples. However, it was not clear which features yielded the highest accuracy for an accent classification task. These five features were used to train a 2-layer CNN on a dataset of five distinct language-accents, namely, Arabic, English, French, Mandarin, and Spanish. The accuracy of each feature were evaluated and compared. The MFCC yielded the highest accuracy.

Keywords: Accent Recognition, MFCC, Chromagram, Spectral Centroid, Spectral Roll-off, Spectrogram.

Accent Recognition is a classification task. Features are representations that are extracted from audio samples, and these features, as images, serve as input to the classification model. There are various features, all of which display frequency information of the speech audio. This work investigated the classification of different features from speech accent audio samples. These features included the Mel-Frequency Cepstral Coefficient (MFCC), Spectrogram, Chromagram, Spectral Centroid, and Spectral Roll-off. The MFCC is most often used in Automatic Speech Recognition (ASR) systems, while the Chromagram is often used for music related tasks. The Spectrogram is sometimes used for ASR related tasks, while the Spectral Centroid and Spectral-roll are not as commonly used. These five features were used to determine if there are features which are competitive to the preferred MFCC feature, as well as identify if accents have distinctive melodies (using the Chromagram).

MFCC features are generally used for accent recognition systems [1], works best for shallow models [2], and represents frequency regions which are audible to the human ear [3]. Spectrograms are representations of speech signal as a sequence of vectors [3]

and performs best with deep models [2]. A 2-layer CNN model was used to classify accents and it achieved an accuracy rate of 77.9% [4]. A fusion of Spectral Centroid features with MFCC features increased Qur'anic accent identification by 4% as compared to using MFCC features alone [5].

The Data was acquired from the Speech Accent Archive [6]. Five language-accents were extracted from the Archive, namely English (627 samples), Spanish (220), Mandarin (132), French (80), and Arabic (172). Speakers recited a standard paragraph. One of the objectives was to identify an optimal segment length for the model to have a more accurate classification. The original dataset was rendered to create 3 different datasets, each of which had speakers reading for different lengths of time such as a single isolated word ("Please"), three consecutive words ("Please call Stella.") and recitation of the paragraph. There were five features investigated, and three datasets, yielding a total of fifteen different experiments. Each Feature was extracted from the raw audio and served as input to a 2-layer CNN model which was used for the classification task.

Table 1. Results for each dataset and each feature combination

Segment Length	Metrics	MFCC	Spectrogram	Chromagram	Spectral Centroid	Spectral Roll-off
Word	**Accuracy**	0.4727	0.4277	0.4727	0.4866	0.4727
	Loss	1.3284	1.6378	1.4079	8.2749	8.4998
3 Words	**Accuracy**	0.5392	0.5060	0.5094	0.5277	0.5288
	Loss	1.3059	1.6599	1.3499	1.3232	7.5943
Paragraph	**Accuracy**	0.4824	0.4590	0.4590	0.4460	0.4727
	Loss	1.3271	1.5432	1.4011	1.4205	8.4998

The MFCC feature was the best performing feature. Three consecutive word utterances performed better than single word utterances or paragraph recitation. The three consecutive word utterances were long enough for the CNN to extract a pattern in the speech, but not too long that a complex pattern appeared, which is not generalizable, as can be seen in the paragraph. English samples made up the majority of the dataset, and was best classified. The MFCC only had 13 bands, and the Chromagram had 12 pitch scales. Therefore, the shallow CNN model was able to identify a pattern between samples using those non-complex features. The Spectrogram is more complex since it records overtones (timbre) and does so by representing lines above each other along its y-axis. Therefore, a shallow model did not grasp a pattern between Spectrograms as well. Although the Spectral Roll-off is not often used in Speech Processing, it achieved the second highest accuracy. The Spectral Centroid, in comparison to the other features' accuracies, was relatively high, and the third best.

References

[1] K. Chakraborty, A. Talele, and S. Upadhya, "Voice Recognition Using MFCC Algorithm," *International Journal of Innovative Research in Advanced Engineering*, vol. 1, no. 10, p. 4, 2014.

[2] Y. Fan, M. Potok, and C. Shroba, "Deep Learning for Audio," University of Illinois at Urbana-Champaign, Department of Computer Science, Illinois, Lecture Proceedings, 2017.

[3] K. Prahallad, "Speech Technology: A Practical Introduction," Carnegie Mellon University & International Institute of Information Technology Hyderabad.

[4] K. Chionh, M. Song, and Y. Yin, "Application of Convolutional Neural Networks in Accent Identification," Project Report, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2018.

[5] N. Kamarudin, S. A. R. Al-Haddad, S. J. Hashim, M. A. Nematollahi, and A. R. B. Hassan, "Feature extraction using Spectral Centroid and Mel Frequency Cepstral Coefficient for Quranic Accent Automatic Identification," in *2014 IEEE Student Conference on Research and Development*, Penang, Malaysia, 2014, pp. 1–6.

[6] George Mason University and S. H. Weinberger, "Speech Accent Archive," 2014. [Online]. Available: http://accent.gmu.edu/.