**Introduction**: (points can also be used in abstract)

The difficulty in accent recognition and speech recognition in general is the problems added with accent specific pronunciation of words. In particular, there are words that have different pronunciations across different accents and on the other hand, there are different words that have the same pronunciation across different accents. For example, words like laugh and class have different pronunciations in US and UK accents. Whereas, different words like floor-flaw and flower-flour has same pronunciation in UK accents, but each of these words has unique pronunciation in Indian accent. We can see that if we know the accent of the speaker before performing speech recognition tasks, it could help in solving some of these problems. But the same problem exists with accent recognition. To solve this problem, we need to find a way to extract features in such a way that there is no dependency of individual word pronunciation in finding the accent. That is where Mel-Frequency Cepstral Coefficients (MFCC for short) becomes really useful. In this paper, we would see how MFCC helps in solving our problem.

**MFCC**:

MFCC is a mathematical trick that converts the input spectrum from the audio file into a vocal signature for each small frame of the input. The mean of the vocal signatures for each of these frames is computed which gives the vocal signature of the speaker. This is used to by most AI based home devices to recognize individuals speaking to them and allows them to have custom behaviors for known users. Our intuition is that speakers of similar accents will have similar vocal signatures compared to speakers of different accents. We would then use this to classify the user’s accent. Below we discuss what MFCC does and how it gives us what we need.

Since the audio signal is constantly changing, we simplify it by dividing it to small frames of 20-40ms where each frame is now relatively constant with minimum changes. Then power spectrum of each frame is calculated which is inspired by human hearing (more specifically cochlea in our ear). This helps in identifying frequencies in the frame. The spectrum still contains a lot of information that we don’t need. The periodogram is then obtained by performing Fourier transformations on the power spectrum.

It is interesting to observe that we cannot distinguish between close frequencies, especially at higher frequencies. For example, if we hear a sound at 100 Hz and 200 Hz, we might notice the difference, but we will probably not be able to differentiate sounds at 1100 Hz and 1200 Hz. We utilize this observation and use logarithm of the periodograms to obtain which allows us to perform cepstral mean subtraction, which is a normalization technique.

There is still overlapping in this cepstral obtained. To handle that, we take the discrete cosine transformation on the logarithmic periodogram to obtain the final Mel-Frequency Cepstral Coefficients. We then compare it with the Mel-scale. We usually keep only 12-26 coefficients as it is observed to degrade the features and based on observation mentioned before, we do not need changes in higher frequencies as they are not noticeable for humans. For our experiments, we have taken only 12 MFCCs.

References:

MFCC - <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/>