Task 1 MFCCs

Baiyang Qu

May 12, 2023

1 Introduction

Speech processing plays an important role in any speech system whether its Automatic Speech Recognition (ASR) or speaker recognition or something else. Mel-Frequency Cepstral Coefficients (MFCCs) were very important feature in the processs of ASR [1]. To get MFCCs features, a signal goes through a pre-emphasis filter; then gets sliced into (overlapping) frames and a window function is applied to each frame; afterwards, we do a Fourier transform on each frame (or more specifically a Short-Time Fourier Transform) and calculate the power spectrum; and subsequently compute the filter banks. To obtain MFCCs, a Discrete Cosine Transform (DCT) is applied to the filter banks retaining a number of the resulting coefficients while the rest are discarded [2]. A final step in both cases, is mean normalization.

2 Baseline

2.1 Implementation process for the baseline

1. Pre-Emphasis—The first step is to apply a pre-emphasis filter on the signal to amplify the high frequencies. To compensate for the high frequency part of the speech signal that is suppressed by the articulatory system by eliminating the effects caused by the vocal folds and lips during the vocal process. It also highlights the high frequency resonance peaks. here, I set the pre-emphasis coefficient = 0.97

2. Framing—After pre-emphasis, I split the signal into short-time frames. The rationale behind this step is that frequencies in a signal change over time, so in most cases it doesn't make sense to do the Fourier transform across the entire signal in that I would lose the frequency contours of the signal over time. Therefore, by doing a Fourier transform over this short-time frame based on the assumption that frequencies in a signal are stationary over a very short period of time, we can obtain a good approximation of the frequency contours of the signal by concatenating adjacent frames. As required, my settings are 25 ms for the frame size, frame size = 0.025 and a 10 ms stride (15 ms overlap), frame stride = 0.01.

3. Window—For smoothing signals, the use of a Hamming window for smoothing reduces the size of the partials and spectral leakage after the FFT.

4. Fourier-Transform and Power Spectrum— I did FFT on each frame to calculate the frequency spectrum, and then compute the power spectrum (periodogram)

5. Filtering with Mel Scale filter banks—Because there is a lot of redundancy in the frequency domain signal, filter banks can be streamlined for the magnitude of the frequency domain, with one value for each band. The amplitude spectrum obtained from the FFT is multiplied and summed with the frequency of each filter, and the value obtained is the energy value of the frame in the corresponding band of the filter. If the number of filters is 22, then 22 energy values should be obtained at this point.

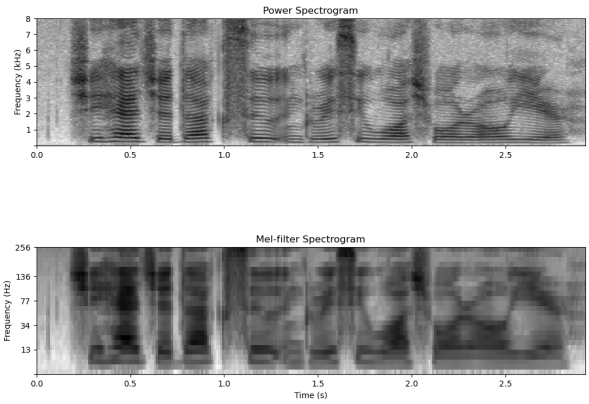
6. Log and DCT—Since the perception of sound by the human ear is not linear, it is better described by the non-linear relationship of the log. According to the definition of the inverse spectrum, this step requires an inverse Fourier transform and then a low-pass filter to obtain the final low-frequency signal. Here the low-frequency information of the frequency spectrum can be

obtained directly using DCT. Since there is an overlap between the filters, the energy values obtained earlier are correlated with each other, and DCT also allows for dimensional reduction and abstraction of the data to obtain the final characteristic parameters.

7. Mean Normalization—As previously mentioned, to balance the spectrum and improve the Signal-to-Noise (SNR), we can simply subtract the mean of each coefficient from all frames.

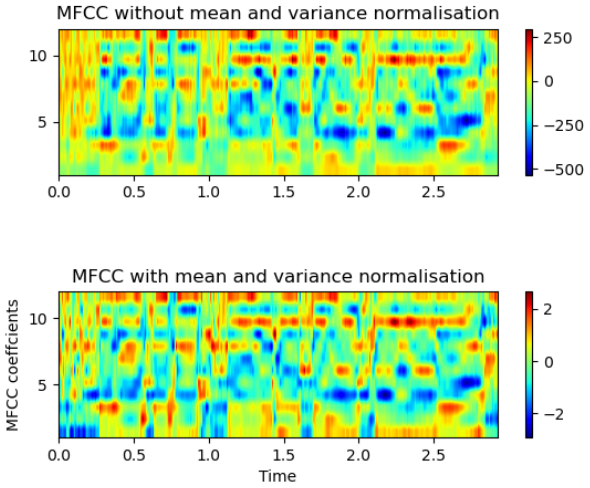
2.2 Baseline results

1. compare mel spectrogram and power spectrogram



As the pictures show, Mel spectrum can be thought of as a "fuzzy" or "compressed" version of the power spectrum. The Mel spectrogram is obtained by applying a filter bank with Mel-spaced frequency bands to the power spectrogram of a signal. The filter bank essentially weights the power spectrogram according to the sensitivity of the human ear to different frequency bands, thus producing a compressed version of the power spectrogram with fewer frequency bins.

2. MFCCs before and after cepstral mean and variance normalization (CMVN)

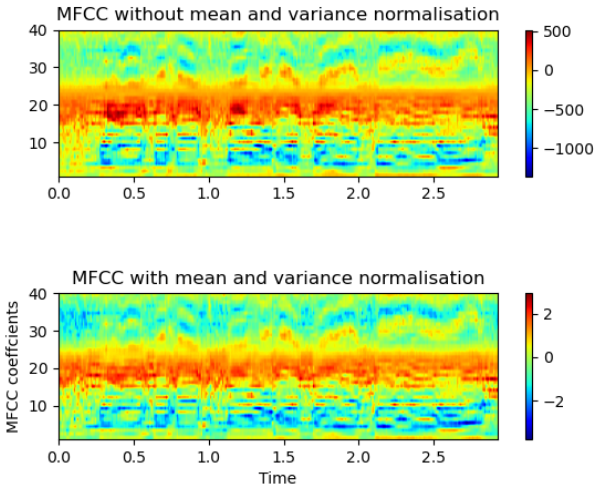


MFCC can be seen as a deeper analysis and processing of the Mel spectrogram, reducing redundant information and extracting more critical features. Prior to cepstral mean and variance normalization (CMVN), the MFCC may have an energy shift, i.e. the component energies of some frequencies are too high or too low. By using CMVN, the MFCC can be normalised to have a zero mean and unit variance. This means that the values in the MFCC are not affected by the energy shift.

3. MFCCs There are vertical bars on the MFCC image where the power spectrogram is as well. When there is a sharp change in the spectral content of the speech signal, such as a sudden onset of a high-frequency sound, this may result in a vertical line in both the MFCC plot and the power spectrogram. However, in general, the vertical lines in the MFCC plot and the power spectrogram may not always correspond directly, as the MFCC computation involves filtering the power spectrum of the speech signal through a set of triangular filters, which can result in smoothing and blurring of the spectral content. The MFCC computation involves filtering the power spectrum of the speech signal through a series of triangular filters that are spaced uniformly on the mel scale. These triangular filters have a bandwidth that increases with frequency, which means that they are wider for higher frequencies. As a result, the MFCC computation tends to emphasize the lower frequency components of the power spectrum more than the higher frequency components.

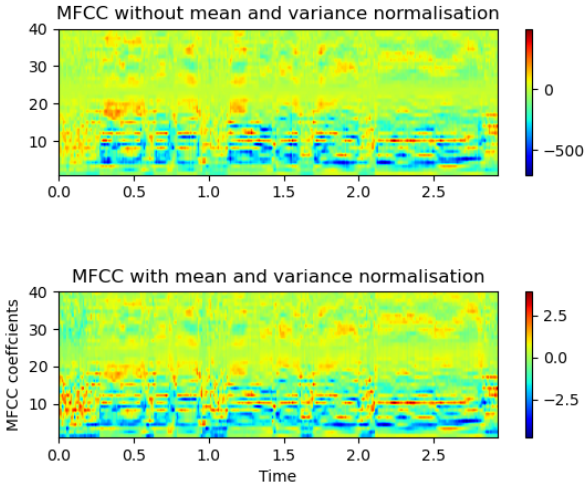
3 other configs and explainations

3.1 80 Filterbanks, 40 MFCCs



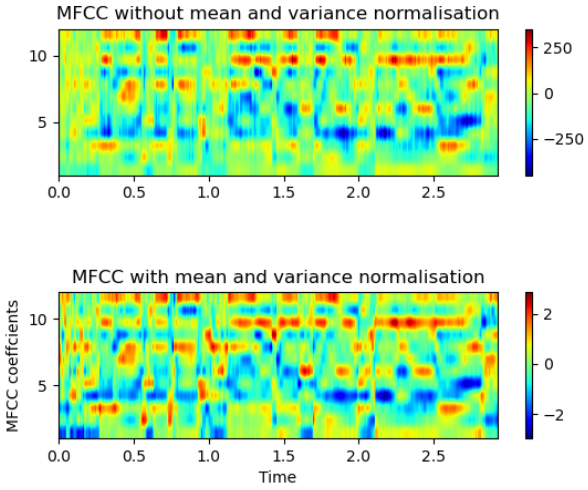
(a) 80 Filterbanks, 40 MFCCs: In this configuration, the number of filterbanks is increased to 80 while keeping the number of MFCCs at 40. The resulting plot of the MFCC features may show a red stripe in the middle of the plot. The red stripe is caused by Spectral leakage phenomenon. It occurs because the Mel-frequency filterbank is not perfectly spaced in the frequency domain. As the number of filterbanks increases, the spacing between the filters becomes narrower and the filterbank becomes more sensitive to small changes in the frequency of the speech signal. This can lead to spectral leakage, where energy from adjacent frequency bands spills over into adjacent filterbanks. The spectral leakage artifact appears as a red stripe in the middle of the MFCC plot because the central frequency of each filterbank is weighted more heavily than the surrounding frequencies in the Mel-frequency scale. This causes a concentration of energy in the central frequency region, which can result in the red stripe.

3.2 40 MFCCs



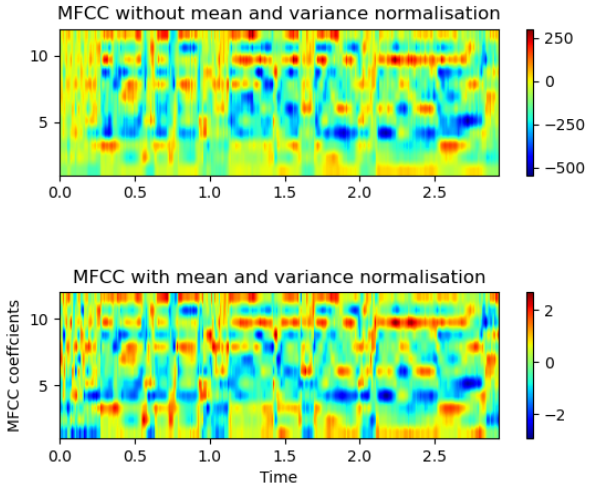
(b) 40 MFCCs: When the number of cepstral coefficients is increased from 12 to 40, the resulting plot of the MFCC features will have more dimensions, which can capture more detailed spectral characteristics of the speech signal. In the resulting plot, we may observe that a wider range of coefficients is evident, with most of the areas with coefficients between 20 and 40 being yellow and green. This pattern is likely due to the fact that the higher cepstral coefficients correspond to higher frequencies in the speech signal. As the number of cepstral coefficients is increased, the plot will capture more information about the higher frequency components of the speech signal. The yellow and green colors in the plot correspond to higher values of the cepstral coefficients, indicating that there is more energy in these higher frequency components.

3.3 No Hamming window



(c) No Hamming window: the MFCCs have a jagged appearance and less clear spectral peaks. In the MFCC images without Cepstral mean and variance normalisation, this is more clearly observed between 2.0 and 2.5. Whereas between 0.0s and 0.3s and part of the image can clearly be seen to change from red to green, representing a decrease in value. The Hamming window is used to reduce spectral leakage in the Fourier transform, which can lead to spurious high-frequency components in the resulting MFCCs. Removing the Hamming window will result in a higher spectral leakage and a less accurate representation of the spectral shape of the signal. As a result, the resulting MFCCs have more high-frequency noise and a less distinct spectral shape.

3.4 No Pre-emphasis.



(d) No Pre-emphasis: The plot has more low-frequency energy and less high-frequency energy compared to the plot produced by the baseline configuration that includes pre-emphasis.

The reason for this difference is that pre-emphasis is a high-pass filter that amplifies the higher frequency components of the speech signal and reduces the low-frequency components. Skipping the pre-emphasis step means that the low-frequency components of the speech signal will not be attenuated, resulting in more energy in these frequencies. Additionally, without pre-emphasis, the high-frequency components will not be amplified, leading to less energy in these frequencies.

In terms of the specific plot of the MFCC features, we may observe that the coefficients corresponding to the low-frequency components of the speech signal will have higher values compared to the baseline plot. At the same time, the coefficients corresponding to the high-frequency components will have lower values. This can be observed visually as a shift in the distribution of energy towards the lower end of the frequency spectrum in the plot.

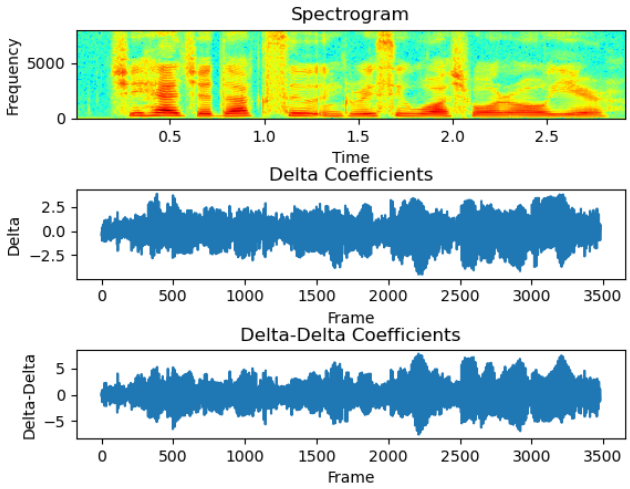
4 First-order and second-order differentials(Extra task)

As the speech signal is continuous in the time domain, the feature information extracted in separate frames only reflects the characteristics of the speech in this frame. In order to make the features more reflective of the time domain continuity, the dimensionality of the information in the front and back frames can be added to the feature dimension. First-order differencing and second-order differencing are commonly used.

```
# Calculation of first and second order difference coefficients
mfcc_feat = feat_mfcc.astype('float32')
mfcc_feat = (mfcc_feat - np.mean(mfcc_feat)) / np.std(mfcc_feat)
mfcc_feat = np.append(mfcc_feat[1:], mfcc_feat[-1:])

delta = sig.convolve(mfcc_feat, [-1, 0, 1], mode='same')
delta_delta = sig.convolve(delta, [-1, 0, 1], mode='same')
```

I visualised the acoustic spectrograms, the differentials separately, and we can observe that the standard representation in modern HMM-based speech recognisers includes static, delta and delta delta coefficients.



Firstly, for the spectrogram, it only reflects the spectral information at each point in time, but lacks information about how the spectrum changes over time. The first- and second-order difference coefficients provide information on the variation of the spectrum over time, allowing the feature matrix to better characterise the dynamics of the speech signal, such as speech rate, intonation and articulation. Secondly, in terms of dimensionality, the addition of first and second order difference coefficients increases the dimensionality of the feature matrix, but in model training, these additional dimensions can provide more information and help improve the classification performance. Therefore, the use of first- and second-order difference coefficients can provide a more comprehensive description of the dynamic characteristics of the speech signal, thus improving speech recognition performance.

References

[1] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22, 2010.

[2] Shikha Gupta, Jafreezal Jaafar, WF Wan Ahmad, and Arpit Bansal. Feature extraction using mfcc. *Signal & Image Processing: An International Journal*, 4(4):101–108, 2013.