Mel-Frequency Cepstral Coefficients Implementation Issues

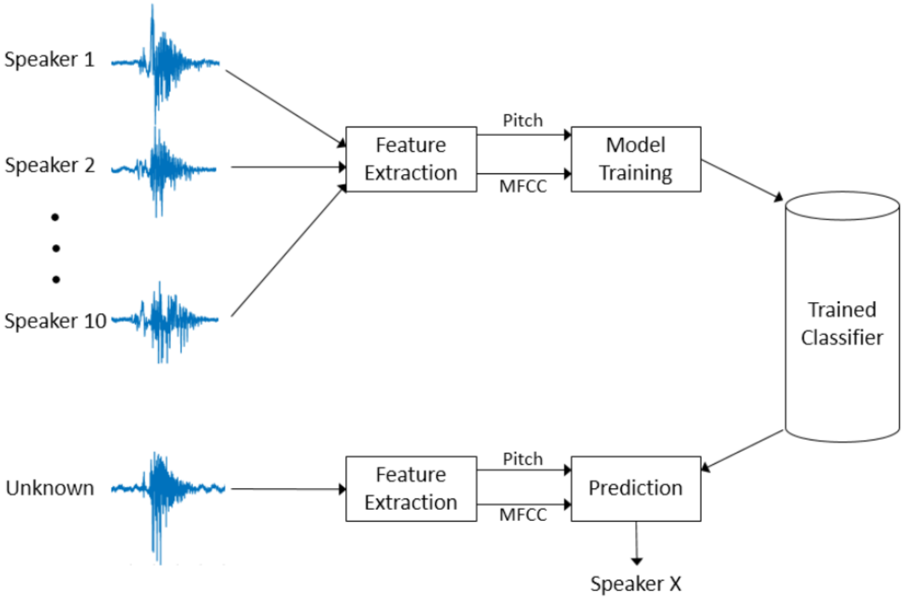
Mel-frequency cepstral coefficients (MFCCs) algorithm is a technique which take voice samples as input and process these signals to calculate coefficients unique to a particular sample. The MFCC algorithm is extensively used in voice recognition systems. Recorded speech signals are samples and stored, and each signal is divided into windows. MFCC is implemented for each of these windows and a set of parameters is extracted per window. Mel filters form a Mel filter Bank, and each filter passes a particular set of frequencies from a frame.

Voice samples of Sid and Sam saying ‘Hello’ at two different instants are passed through the MFCC algorithm and their respective MFCC Coefficients are extracted. Two samples are collected per speaker since one is stored as a template in the database and one is real time input. Then, the two MFCCs are compared for both Sid and Sam. Later, the Euclidean distance is calculated to compare the template and the real time input. This is how MFCC was used to recognize the voices of Sid and Sam.

The following are implementation issues of the Mel-Frequency Cepstral Coefficients algorithm:

- MFCCs are not very robust in the presence of noise due to its dependence on the spectral form. All MFCC are altered by the noise signal even if one frequency band is distorted. For example, if a window was open and Sid and Sam’s recordings also had the traffic noises outside, then this would affect the performance of the MFCC algorithm.
- The performance is affected by the number of filters. The rate of

accuracy fluctuates to increase or decrease the number of MFCC coefficients (Singh, Khan, Shree, 2012). If the number of filter banks is increased, the distortion measure will increase, since there is an increase in the number of terms in Euclidean distance. If Sid and Sam used too many or too little filters, this would affect the accuracy of the MFCC algorithm.

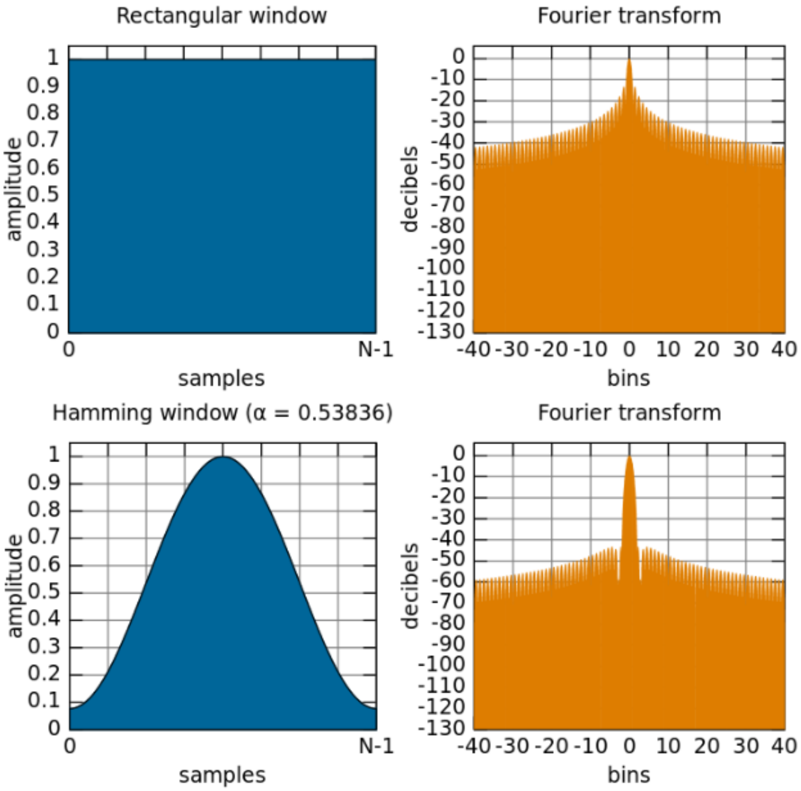


The above figure shows how MFCC is used for speaker identification for the unknown speaker after being trained on 10 speakers. Pitch and MFCC are extracted from the speakers.

- The Filterbank design affects the robustness of MFCC. A certain shape of Filterbank is required but this is impractical since it needs re-tuning, change of environment, and selection of filter shape at each use.
- Filter bandwidth is not an independent design parameter, but instead is calculated by the filter bank and the number of filters used in recognition (Singh, Khan, Shree, 2012).
- MFCCs are sensitive to channel mismatch when undergoing

training and testing. The type of channel used for creating speaker model is critical for good performance of a speaker recognition system (Chougule, Chavan, 2014).

- Mismatch between training and testing models gravely affects the speaker identification accuracy. An example of vocal mode mismatch is when a system is trained with data acquired under laboratory conditions while the test data is acquired under real environments with different active noise sources (Sadjadi, Hansen).



The figure above depicts all the frequencies being subjected to the Fourier transform of the window function. The top graph of the diagram is due to a Rectangular window and the bottom graph is due to a Hamming window

- The voiced part of the speech signal consists of pertinent information about speaker identification. If the redundant information is not removed, during preprocessing, from the unvoiced part, the processing signal will be too large and take more time and memory to extract features from the signal. Hence, we have to go through another step to ensure the training of the neural network does not take too long.

- MFCCs cannot represent high frequency ranges as well as they do low frequency ranges. This is due to larger spacing of filters in the high frequency range.
- The Filterbank and Mel-frequency warping can be implemented by computing a Mel-warped spectrum by interpolation from the original discrete-frequency power spectrum. But, this results in the discretization being especially critical due to large, dynamic range of the power spectrum (Molau, Pitz, Schluter, Ney).

Khan, R. Shree, R. Singh, N. MFCC and Prosodic Feature Extraction Techniques: A Comparative Study. (2012, September). Retrieved from: https://pdfs.semanticscholar.org

Chougule, S. Chavan, M. Advances in Intelligent Systems and Computing. Retrieved from: https://link.springer.com

Sadjadi, S. Hansen, J. Assessment of Single-Channel Speech Enhancement Techniques for Speaker Identification under Mismatched Conditions. Retrieved from: https://pdfs.semanticscholar.org

Molau, S. Pitz, M. Schluter, R. Ney, H. Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum. Retrieved from: http://kom.aau.dk