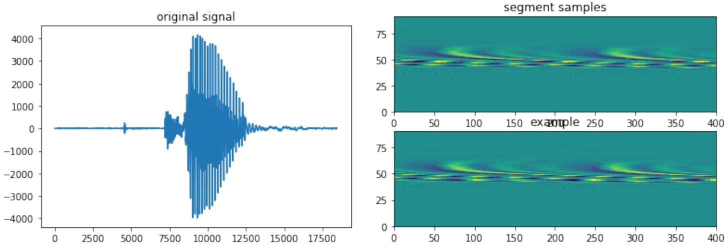
KTH ROYAL INSTITUTE
OF TECHNOLOGY

DT2119 Speech and Speaker Recognition Lab 1

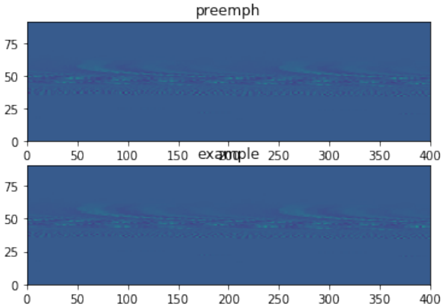
Zijian Fan, Lingxi Xiong

MFCC – Enframe



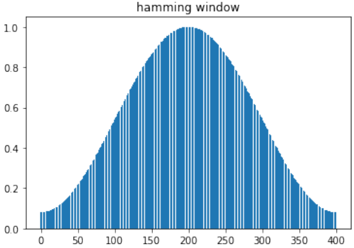
- window length = 400(200ms); window shift = 200(100ms)–50% overlap
- increase the resolution and decrease the variance

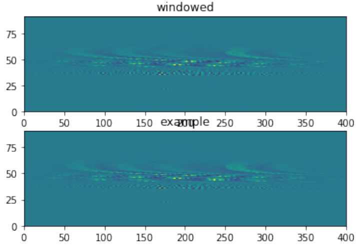
MFCC – Pre-emphasis



- high-pass filter to filter out the low frequency component
- $H_{pre}(z) = 1 - 0.97z^{-1}$

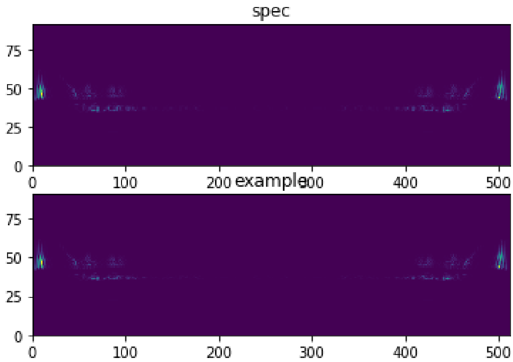
MFCC – Hamming Window





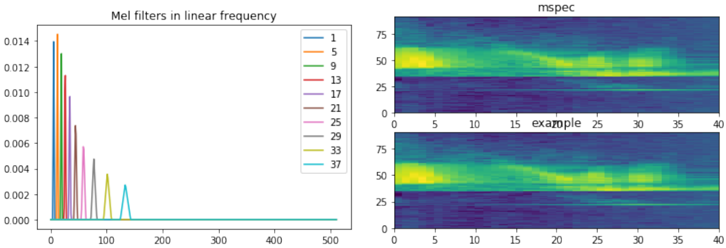
- $h[n] = h_l[n] \times w[n]$
- Resolution: $\Delta\upsilon = 1.30/N$
- Variance: $Var\{\widehat{P_x^W}(v)\} = \frac{9}{8K} P_x^2(v) = \frac{9}{8K} Var\{\widehat{P_x}(v)\}$

MFCC – Fast Fourier Transform



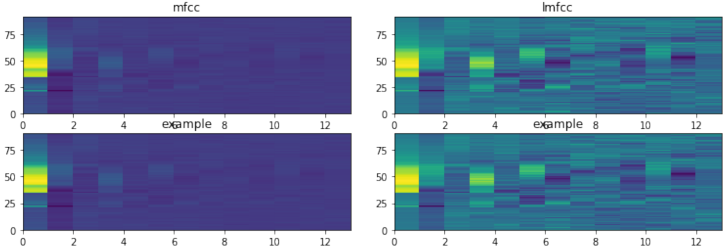
- f_{max} = 10000Hz

MFCC – Mel filterbank log spectrum



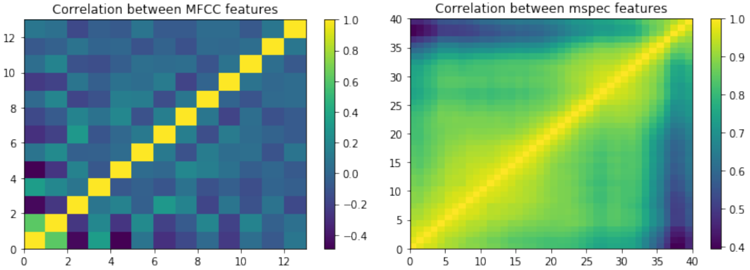
- Mel filterbank: filters concentrated in the low frequency area

MFCC – Cosine Transform and Liftering



- Cosine transform: continous → discrete signal
- Lifter: correct the range of the coefficients
- the MFCCs are similar if from same digits by same speaker, otherwise they vary a lot

Feature Correlation



- MFCC: uncorrelated features → diagonal covariance matrices → Gaussian modelling
- Mspec: features are much more correlated

Explore Speech Segments with Clustering

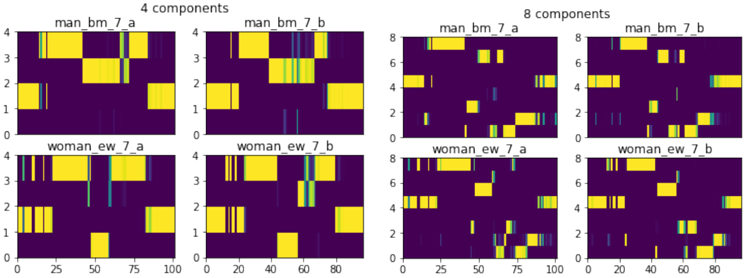
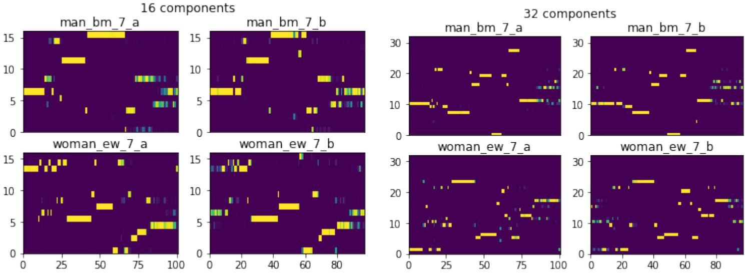


Figure: GMM posteriors of utterances containing same words

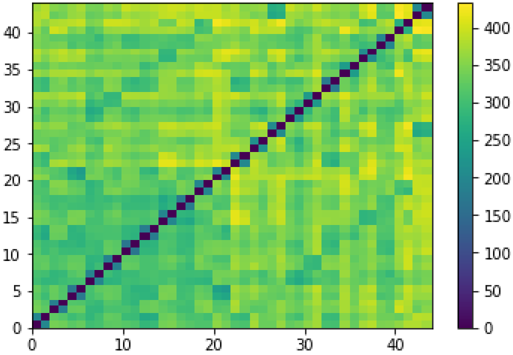
- the discovered classes increase with number of components increase

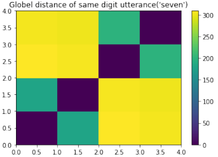
Explore Speech Segments with Clustering



- The classes does not correspond to the phonemes composing each word
- Unstable: classes that represent the utterances(word) vary among speakers

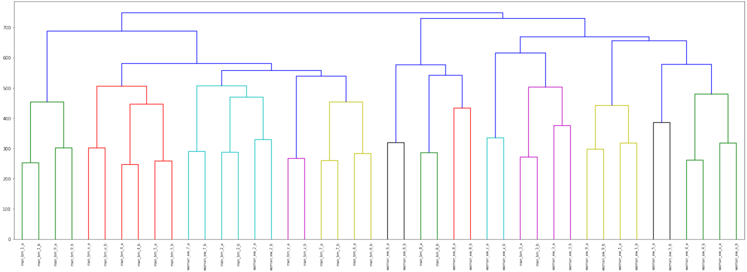
Comparing Utterances – Global distance




Globel distance of same digit utterance('seven')

- The distance separates digits well even between different speakers

Comparing Utterances – hierarchy clustering



Thank you for your Attention!