# Mel-frequency cepstral coefficients (MFCCs) and Dynamic Time Warping (DTW) based Automatic Speech Recognition Algorithm


Developed by: Roberto Costa

Teachers: Federico Avanzini, Giovanni De Poli


Date: 29 / 11 / 2017

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

# Objectives

- Developement of a voice commands recognition algorithm which can be trained on a dataset
  with the purpose of recognizing different commands and different speakers.

- Verify the affidaility of the algorithm by a K-fold cross validation.

# Algorithm

- Data acquisition
  fc = 8 kHz
- Mel-frequency cepstral coefficients computation
  Hamming window in time domain
  Triangular shaped filters in mel domain
  # of filters in filterbank (approx. 2.1 per octave) : floor(3*log(fc)) = 11
  filters act in the absolute magnitude domain
  highest filter (0.5 fc) taper down to zero
  overlap: 50%
  length of frame in samples: $n = \min_i \left\{ i : 2^i < \frac{3}{100} f_c \right\}$

http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/melcepst.html

# Algorithm

- Data splitting between training and testing
- Dinamic Time Warping algorithm between every couple of MFCCs sequences in the training dataset.
  Each MFCCs sequence is a vector of elements composed by 12 coefficients.
  The DTW algorithm computes the cost matrix for each couple of vectors, given by the Euclidean distance between elements;
  then finds the optimal path and returns the Euclidean distance between mapped couples of elements.
  CONFIDENCE INDEX DYNAMIC TIME WARPING FOR LANGUAGE-INDEPENDENT EMBEDDED SPEECH RECOGNITION - Xianglilan Zhang, Jiping Sunz, Zhigang Luo, Ming Liy - IEEE 978-1-4799-0356-6/13 ©2013
- Euclidean distance based clustering of MFCC vectors for each class in C=2 sub-classes, with K-means algorithm.
  C has been choosen to be equal to the number of speakers of the dataset
- For each voice command and for each speaker, a median MFCC sequence has been computed with MFCCmultipleMean function, explained in the next slide

# Algorithm - MFCCmultipleMean

1. Input: set of MFCC sequences
2. Taking the two closest MFCC sequences from the input set
3. Compute the mean with the mapping algorithm given from DTW
4. Delete the couple of MFCC seq. from the input set
5. Adding the mean to a different set
6. Repeating from step 2. until the input set has only a couple
7. Output set: X(i), with $i \in \{1, \cdots, N\}$
8. Y(1) = X(1)
   Y(n) = 0.9 Y(n-1) + 0.1 X(n)
9. Output: Y(N)

# Algorithm

- Computing mean DTW distance and variance of the distance from the median MFCC to add soft information (reliability of the median MFCC)
- Class prediction – version 1:
  Taking the label of the closest median MFCC (the smaller DTW distance)
- Class prediction – version2:
  Predict2 function
  Compute the distance *d_{1,i}, d_{2,i}* with *i* in {*1, .. , # of commands*} between the 2 sub-classes of speakers and the tested vector
  Compute the distance *d_{3,i}, d_{4,i}* beween the tested vector and the mean vector between the closest vector for the first [/second] speaker (sub-class 1 [/2]) and the same labeled vector of speaker 2 [/1]
  Minimize the sum of the distances for each of the two possible labels:
  argmin ( [*d_{1,i} + d_{3,i}, d_{2,i} + d_{4,i}* ] )

# Data validation

**K-fold cross validation**

- The dataset has been split in K={5, 10} parts.
- K-1 parts have been used for training
- 1 part has been used for testing.

# Results

- The mean accuracy is 97.65 %

# Future work

- Frequency response inversion of microphone for recording the training dataset
- Recognition of the speaker