Each frame of utterance : n^{th} frame

$$c_0[n],\ c_1,\ c_2,\ \ldots\ c_{12}[n] \leftarrow 13$$

$$\Delta_1[n] = \frac{c_1[n+1] - c_1[n-1]}{2}$$

$$\Delta\Delta_1[n] = \frac{\Delta_1[n+1] - \Delta_1[n-1]}{2}$$

→ 39-element feature vector.

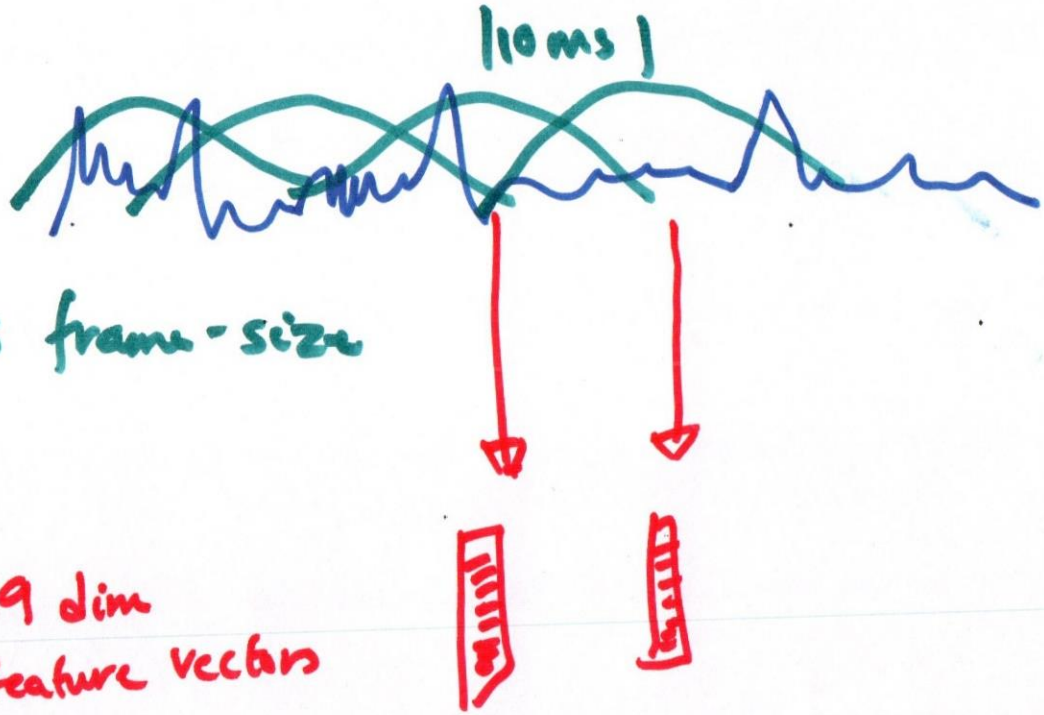
Distance measure: for a test frame w.r.t. ref frame

Cepstral dist. $d^2_{cep} = \sum_m (r_m - t_m)^2$

$m \leftarrow$ cepstral index

$= \sum_m [20 \log(R_m) - 20 \log(T_m)]^2$ ← log spec. dist.

t : test frame
r : ref "



Digit Recognition task

10 CBs, one for each of 0, zero, one, two, ...

We create a "Codebook" CB for each digit containing all the feature vectors of the reference utt.

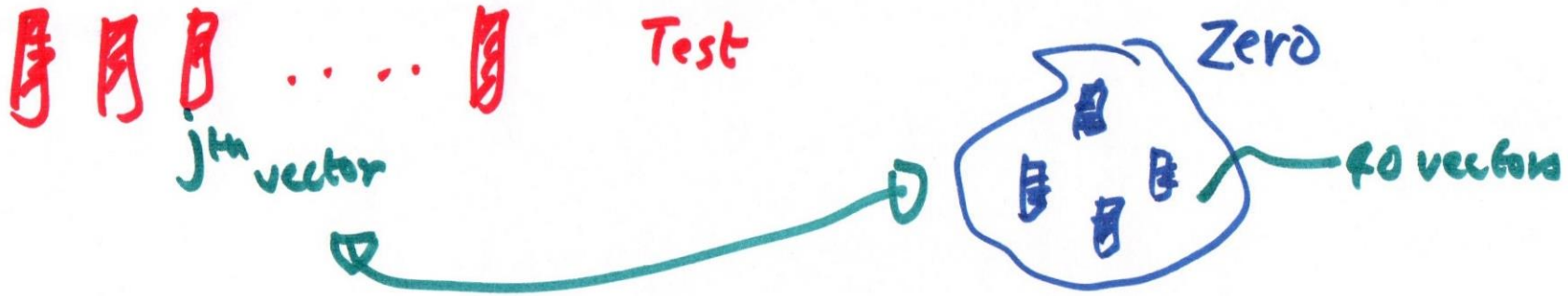
We find the CB that provides the lowest avg. distortn based on the NN rule

$$C = \arg\min_{k} \sum_{j=1}^{N} d_{kj}$$

N ← # test frames

d_{kj} = min distortⁿ of the j^{th} test vector w.r.t. k^{th} CB.

Eucl dist of 39-dim vectors



bin
nib

We need to consider the spectral variability in speech:

- physiology, style, accent, dialect
- phonetic context, prosody, emotion, pace

Vector Quantiz^n (VQ)

• "Codebook" Training

↓

created by a clustering algorithm to get the selected # classes.

Each class is represented by a prototypical vector ("codeword")

("39" dim)

Algo: "K-means"

The algo divides the training set vectors $\{\bar{x}\}$ into $K = L$ clusters C_i s.t. 2 conditions hold:

(i) The codeword $\bar{y}_i$ is chosen to minimise the avg. distortion in the cell C_i

i.e. $\bar{y}_i$ is the vector that minimizes

$$D_i = \mathcal{E}\{ d(\bar{x}, \bar{y}) \mid \bar{x} \in c_i \}$$

$$= \frac{1}{M_i} \sum_{\bar{x} \in c_i} d(\bar{x}, \bar{y}) \quad , \quad \bar{x} \in c_i$$

$$\Rightarrow \bar{y}_i = \text{"centroid"}\ (c_i)$$

(ii) Given the codevectors, $\bar{y}_i$, the training set vectors are assigned to clusters based on the NN rule.

i.e. choose c_i for $\bar{x}$ if it minimizes $d(\bar{x}, \bar{y}_i)$

Iterative Algorithm for Clustering

1. Initialise the CB by choosing L code vectors (randomly from train set)

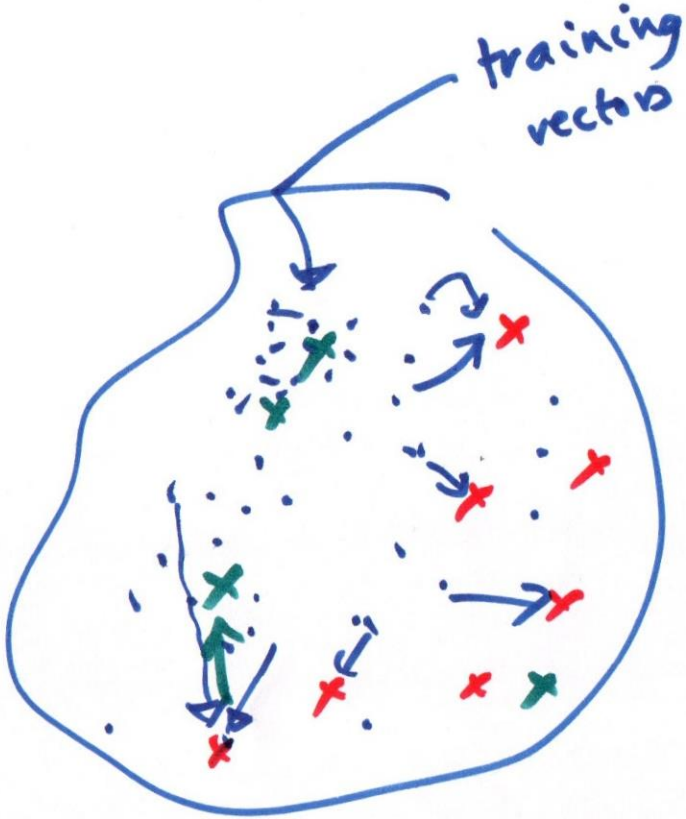
CDEEP
IIT Bombay

2. Classify the training set vectors into L clusters using NN rule for each vector.

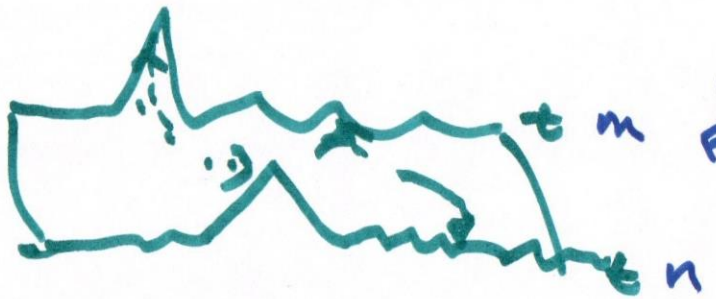
3. Update the codevector of each cluster by the centroid of the training vectors of the cluster.

4. Compute the total distortion over the training set vectors. Stop if it is acceptable. Else go to Step 2.

Repeat choosing different sets of initial vectors to find Global optimum.



DTW



$R(n)$, $T(n)$ ← ref & test patterns (seq of frames of 39-dim feat. vectors)

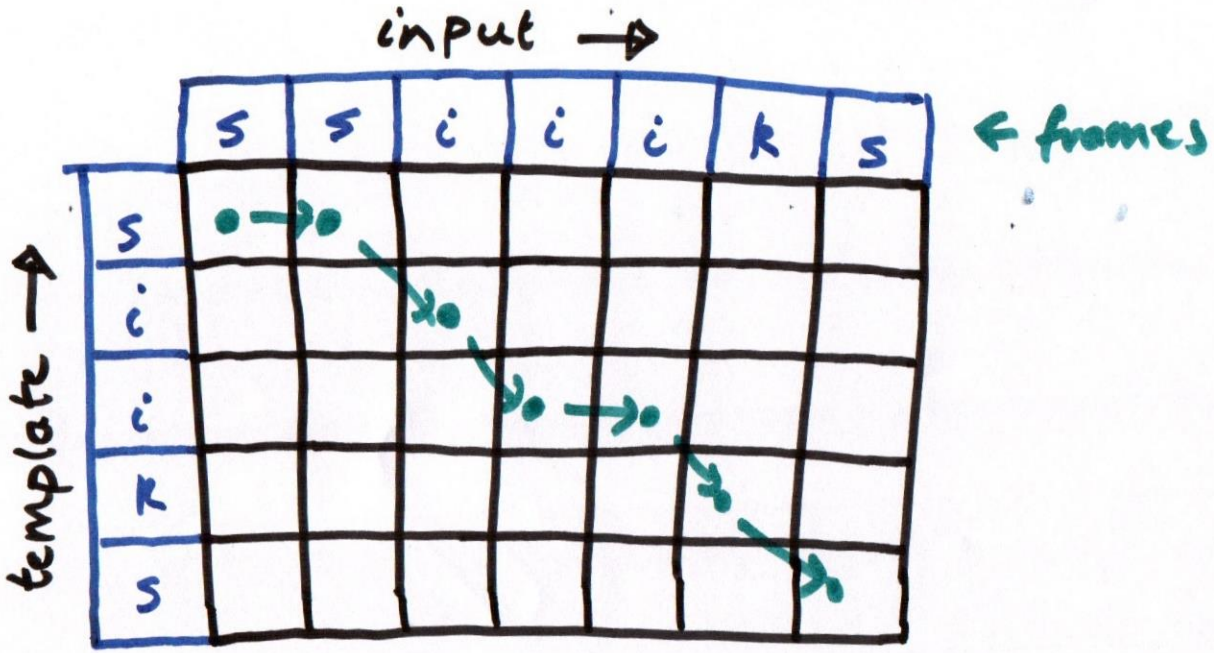
DTW finds a warping fn $m = w(n)$ optimizn.:

$$D = \min_{w(n)} \left[\sum_{n=1}^{T} d\left(T(n), R(w(n))\right) \right]$$

"six"

input →

↓ template



$M \times N \times 3$

$\sim N^2 \quad O(N^2)$

$$D_{m,n} = d_{m,n} + \min\{D_{m,n-1}, D_{m-1,n-1}, D_{m-2,n-1}\}$$

$d_{m,n}$: cepstral dist betw 2 frames.