# A New Clustering technique for speech samples

Sunil Kopparapu, Meghna Maiti, Akhilesh Srivastava, P.V. S. Rao

Cognitive Systems Research Laboratory, Tata Infotech Limited,
Plot No 14, Sector 24, Vashi-Turbe, Navi Mumbai 400 705, India.

**Abstract**

Clustering of data is a well addressed problem in several fields and finds use in various applications. Clustering of speech samples is important in several applications. One of the uses of clustering is in deterministic (non-HMM) based speech recogntion, where selection of ideals is a major issues. In this paper, we propose a novel clustering technique that aids in determining the ideals for a set of speech data. The idea is to use a FoM (Figure of Merit) measure and identify clusters. The speech sample corresponding to the centroid of each cluster represents the cluster and is the ideal. These ideals then are used as references to recognise the test samples. Several experiments with simulated and real data are done to determine the usability of the proposed clustering scheme. La updated 28/11/2005 (16:48)

## 1 Introduction

## 2 Method for Forming Clusters

Consider, for example $N$ speech samples of the same spoken word which needs to be clustered into $n$ clusters.

1. Initialization (divided into groups of $n$ samples in each cluster resulting in $\left[\frac{N}{n}\right]$ clusters)
   (a) They can be divided into groups of $n$ samples in each cluster resulting in $\left[\frac{N}{n}\right]$ clusters.
2. compute the DTW (dynamic time warping) cost of each of the $N$ speech samples with all other $N-1$ speech samples. There will be a total of $\frac{N\times(N-1)}{2} - N$ comparisions[1] costs associated with these $N$ speech samples. The DTW cost represents the cost incurred to warp one speech sample to another.
3. In each cluster find the average of only the first $n-1$ pairs of comparisons for each sample as we do not consider comparison of the sample with itself
4. 5. Repeat step 4 for all the samples (eg: S2,S3,S4,S5.....S30), and get the average score of first 4 pairs of comparisons for each individual samples.

[1]Neglecting comparison of each sample with itself as it is zero results in $-N$

5. 6. Find the minimum average score from all the averages. (eg: Consider that S5 gives the minimum average score for all the samples from S1 to S30).

6. 7. Consider the first 4 comparisons of the sample giving the minimum average (eg: here S5), they form the first cluster C1 which has 5 elements including S5 as an element of the group. The first element of the cluster (eg: here S5) will be the centroid of the first cluster C1.

7. 8. Remove these 5 samples from the initial 30 samples, which form first cluster C1.

8. 9. Again find the average of first 4 pairs of comparison for the remaining (30-5=25) samples.

9. 10. Repeat from step 4 till we are done with all the samples.

10. 11. In this case we get 6 clusters C1, C2, C3, C4, C5, C6, and each cluster has a centroid which can be taken as the reference for that cluster.(eg: in C1, S5 is the reference).

11. 12. After finding the references from each clusters, again compare the remaining samples with these references (in this case 6 references and 24 remaining samples, 30-6=24) excluding the reference samples.

### 2.1 Figure Of Merit (FOM)

1. Find the minimum, maximum and average values for each cluster. 2. Compare the minimum of minimum of intra clusters with minimum of inter clusters. Min(Min of intra clusters) Min of inter clusters 3. Similarly compare the maximum value and average value between intra clusters and inter clusters.

Observation 1: The Figure Of Merit for the initially formed clusters, by dividing equal number of elements in each group will always be larger than the FOM formed by the final group after figuring the centroids and comparing each sample with the centroids and generating the groups.

Observation 2: The FOM values help us to identify the goodness of the clusters formed. The closer to zero the values of FOM are the better are the clusters formed.

Below are examples of the clusters formed, the FOM for 6 clusters is better than the FOM for 3 clusters.

Fig 1: Plot and FOM values for 6 Clusters

[pic]

Fig 2: Plot and FOM values for 3 Clusters

[pic]