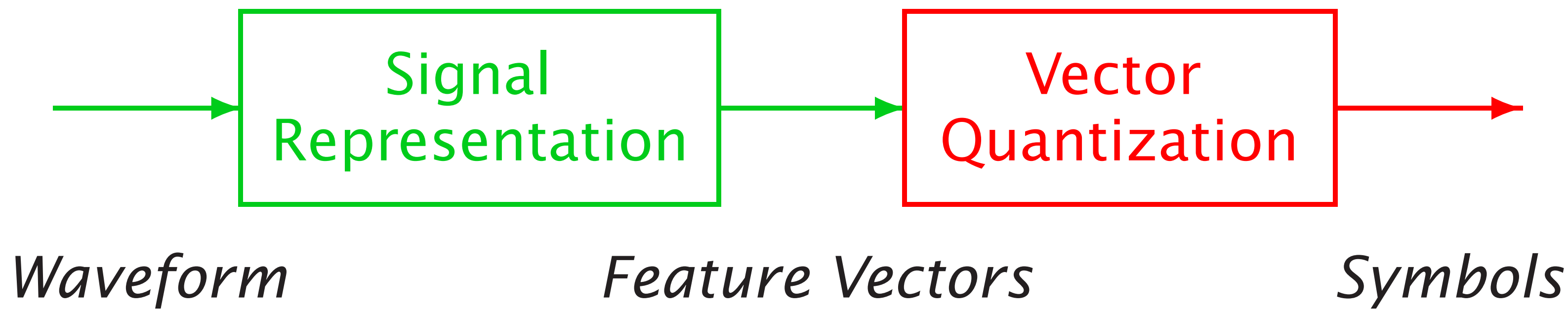
Vector Quantization and Clustering

- Introduction
- K-means clustering
- Clustering issues
- Hierarchical clustering
  - Divisive (top-down) clustering
  - Agglomerative (bottom-up) clustering
- Applications to speech recognition

Acoustic Modelling

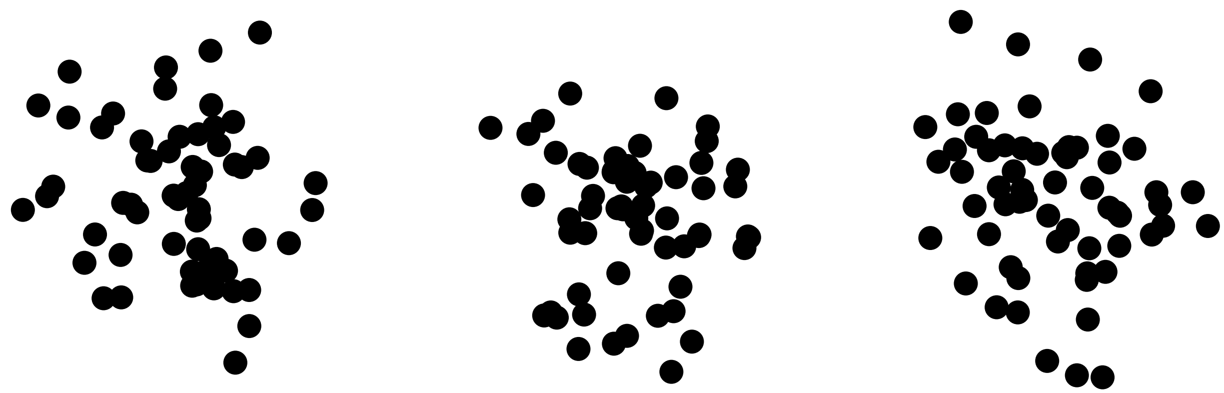


- Signal representation produces feature vector sequence
- Multi-dimensional sequence can be processed by:
  - Methods that directly model continuous space
  - *Quantizing* and modelling of discrete symbols
- Main advantages and disadvantages of quantization:
  - Reduced storage and computation costs
  - Potential loss of information due to quantization

Vector Quantization (VQ)

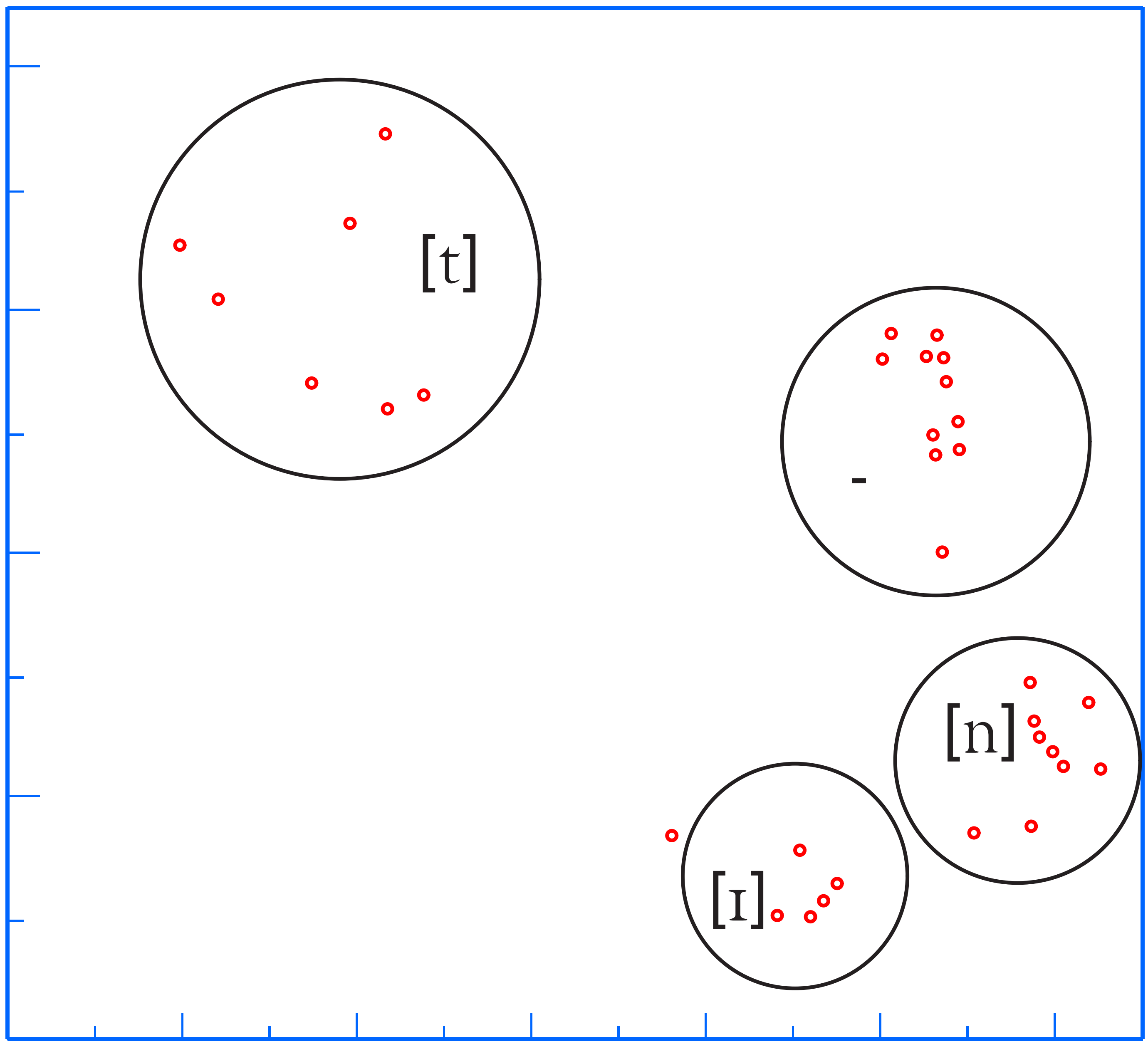
- Used in signal compression, speech and image coding
- More efficient information transmission than scalar quantization (can achieve less that 1 bit/parameter)
- Used for discrete acoustic modelling since early 1980s
- Based on standard clustering algorithms:
  - Individual cluster centroids are called codewords
  - Set of cluster centroids is called a codebook
  - Basic VQ is K-means clustering
  - Binary VQ is a form of top-down clustering (used for efficient quantization)

VQ & Clustering



- Clustering is an example of unsupervised learning
  - Number and form of classes $\{C_i\}$ unknown
  - Available data samples $\{\boldsymbol{x}_i\}$ are unlabeled
  - Useful for discovery of data structure before classification or tuning or adaptation of classifiers
- Results strongly depend on the clustering algorithm

Acoustic Modelling Example



Clustering Issues

- What defines a cluster?
  - Is there a prototype representing each cluster?
- What defines membership in a cluster?
  - What is the distance metric, $d(\boldsymbol{x}, \boldsymbol{y})$?
- How many clusters are there?
  - Is the number of clusters picked before clustering?
- How well do the clusters represent unseen data?
  - How is a new data point assigned to a cluster?

K-Means Clustering

- Used to group data into K clusters, $\{C_1, \ldots, C_K\}$
- Each cluster is represented by mean of assigned data
- Iterative algorithm converges to a local optimum:
  - Select K initial cluster means, $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$
  - Iterate until stopping criterion is satisfied:
    1. Assign each data sample to the closest cluster

       $$\boldsymbol{x} \epsilon C_i, \quad d(\boldsymbol{x}, \boldsymbol{\mu}_i) \le d(\boldsymbol{x}, \boldsymbol{\mu}_j), \quad \forall i \neq j$$

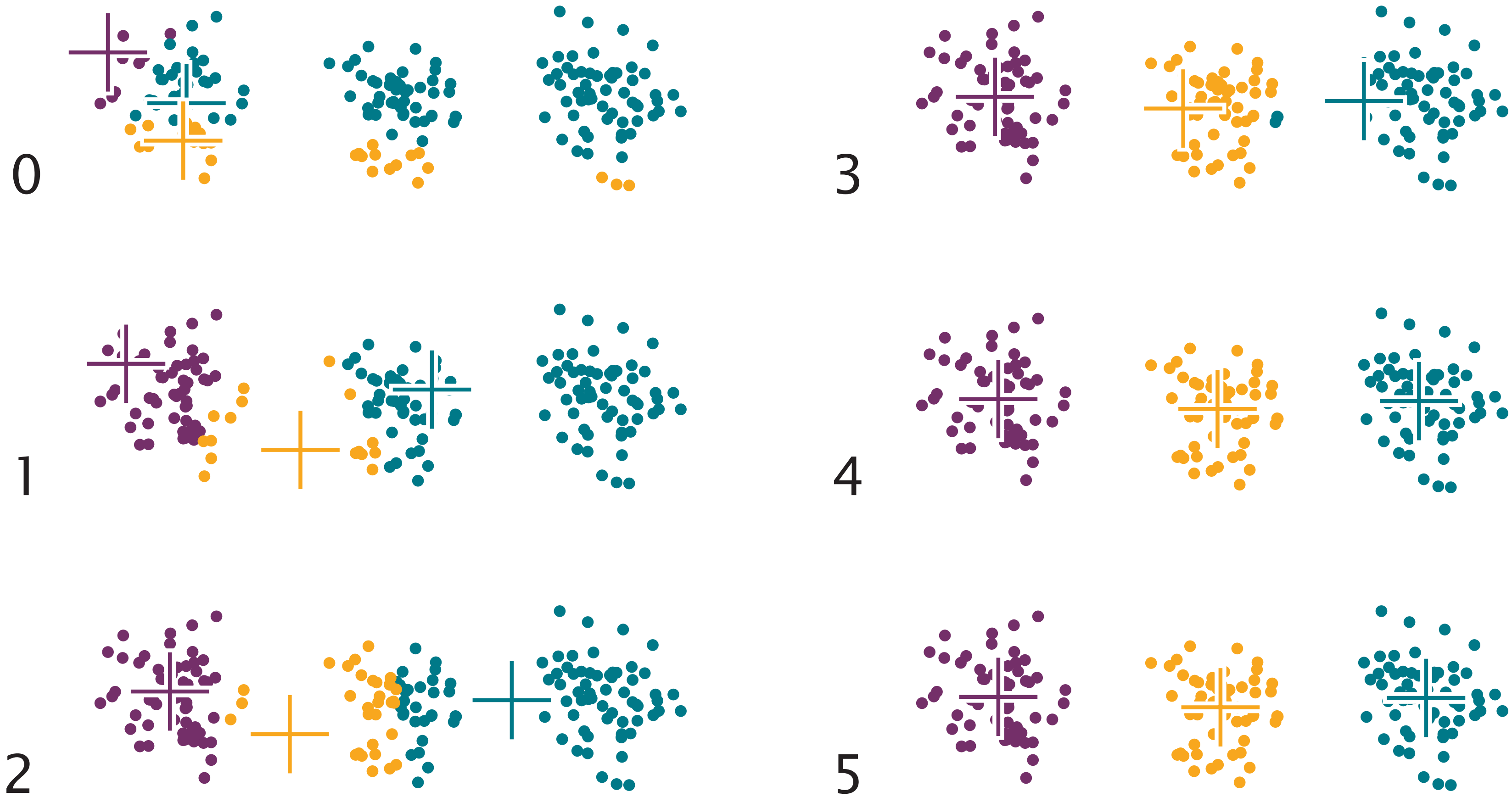
    2. Update K means from assigned samples

       $$\boldsymbol{\mu}_i = E(\boldsymbol{x}), \quad \boldsymbol{x} \epsilon C_i, \quad 1 \le i \le K$$

- Nearest neighbor quantizer used for unseen data

K-Means Example: *K* = 3

- Random selection of 3 data samples for initial means
- Euclidean distance metric between means and samples



K-Means Properties

- Usually used with a Euclidean distance metric

$$d(\boldsymbol{x}, \boldsymbol{\mu}_i) = \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 = (\boldsymbol{x} - \boldsymbol{\mu}_i)^t(\boldsymbol{x} - \boldsymbol{\mu}_i)$$

- The total distortion, $\mathcal{D}$, is the sum of squared error

$$\mathcal{D} = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \epsilon C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

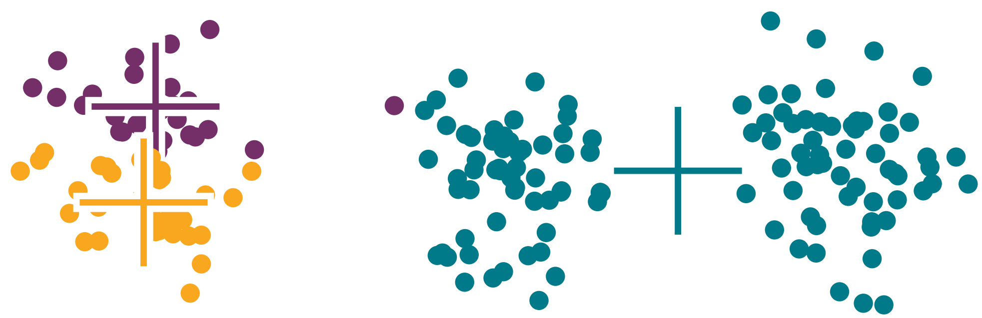
- $\mathcal{D}$ decreases between n^{th} and $n+1^{st}$ iteration

$$\mathcal{D}(n+1) \leq \mathcal{D}(n)$$

- Also known as Isodata, or generalized Lloyd algorithm
- Similarities with Expectation-Maximization (EM) algorithm for learning parameters from unlabeled data

K-Means Clustering: Initialization

- K-means converges to a local optimum
  - Global optimum is not guaranteed
  - Initial choices can influence final result



$K = 3$

- Initial K-means can be chosen randomly
  - Clustering can be repeated multiple times
- Hierarchical strategies often used to seed clusters
  - Top-down (divisive) (e.g., binary VQ)
  - Bottom-up (agglomerative)

K-Means Clustering: Stopping Criterion

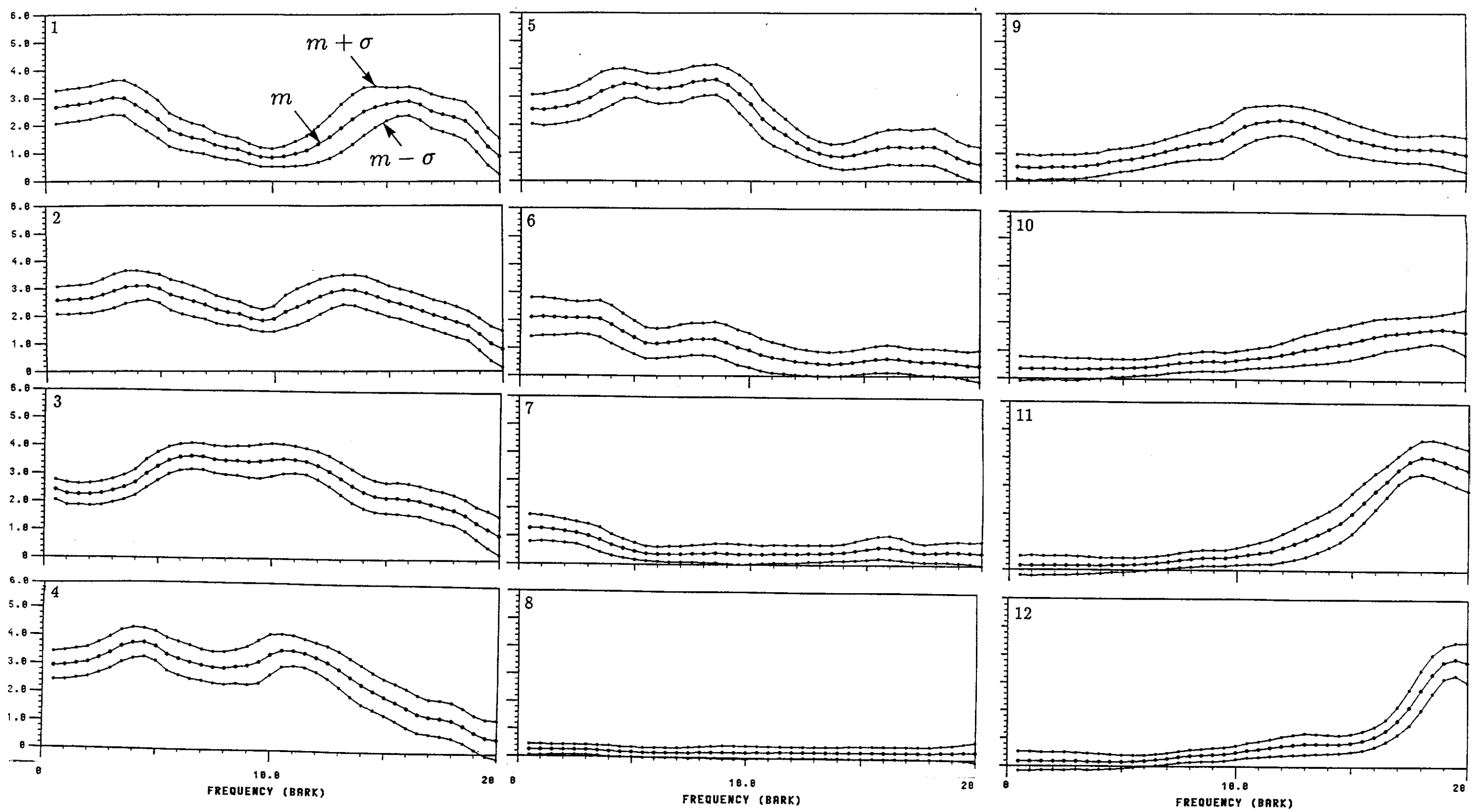
Many criterion can be used to terminate K-means :

- No changes in sample assignments
- Maximum number of iterations exceeded
- Change in total distortion, $\mathcal{D}$, falls below a threshold

$$1 - \frac{\mathcal{D}(n+1)}{\mathcal{D}(n)} < T$$

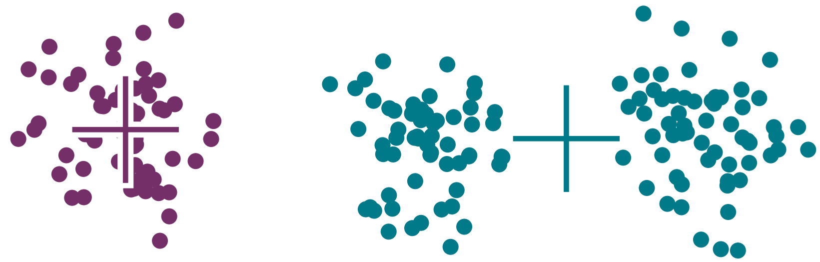
Acoustic Clustering Example

- 12 clusters, seeded with agglomerative clustering
- Spectral representation based on auditory-model

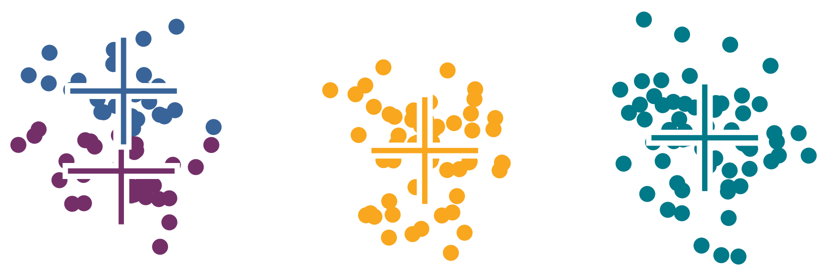


Clustering Issues: Number of Clusters

- In general, the number of clusters is unknown



$K = 2$

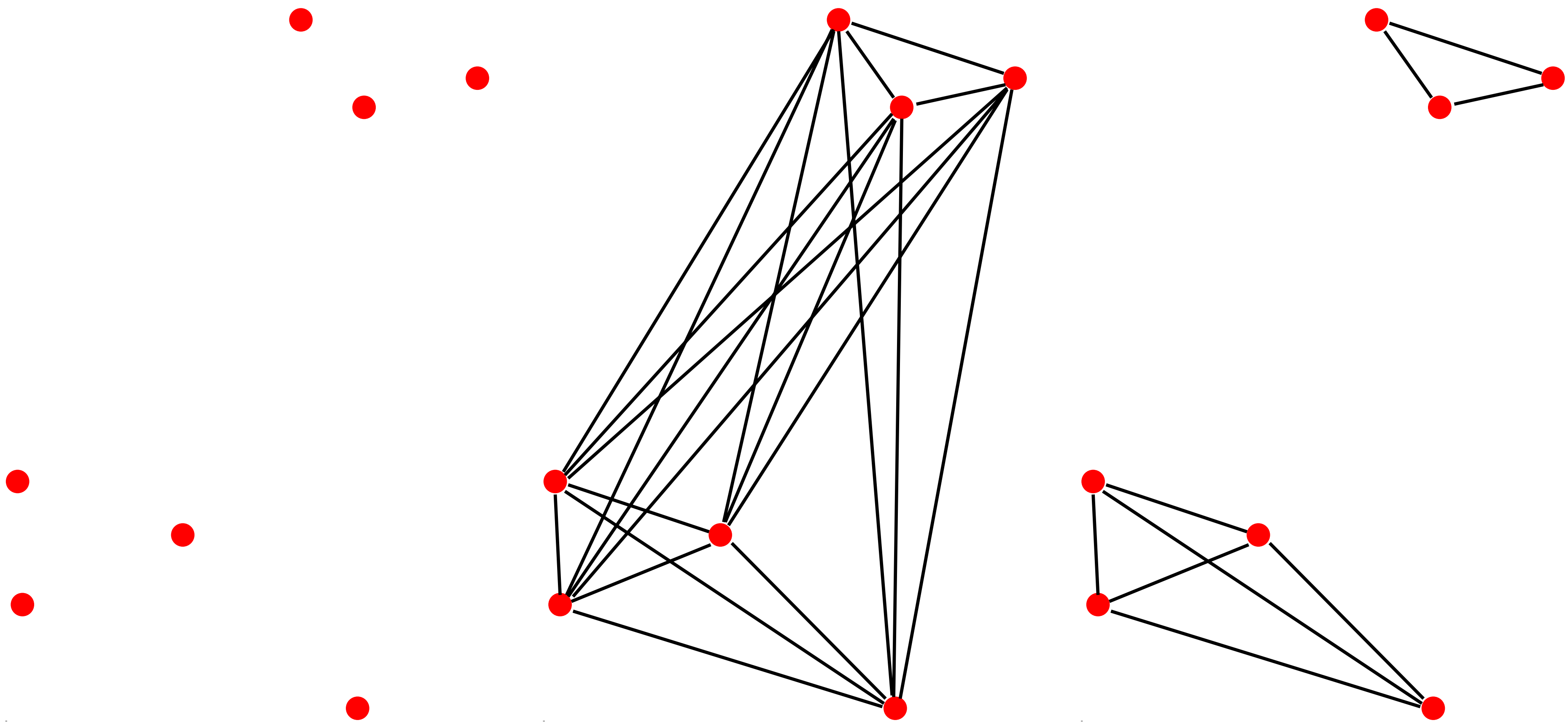


$K = 4$

- Dependent on clustering criterion, space, computation or distortion requirements, or on recognition metric

Clustering Issues: Clustering Criterion

The criterion used to partition data into clusters plays a strong role in determining the final results



Clustering Issues: Distance Metrics

- A distance metric usually has the properties:
  1. $0 \leq d(\boldsymbol{x}, \boldsymbol{y}) \leq \infty$
  2. $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ iff $\boldsymbol{x} = \boldsymbol{y}$
  3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$
  4. $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{y}, \boldsymbol{z})$
  5. $d(\boldsymbol{x} + \boldsymbol{z}, \boldsymbol{y} + \boldsymbol{z}) = d(\boldsymbol{x}, \boldsymbol{y})$ (invariant)
- In practice, distance metrics may not obey some of these properties but are a measure of dissimilarity

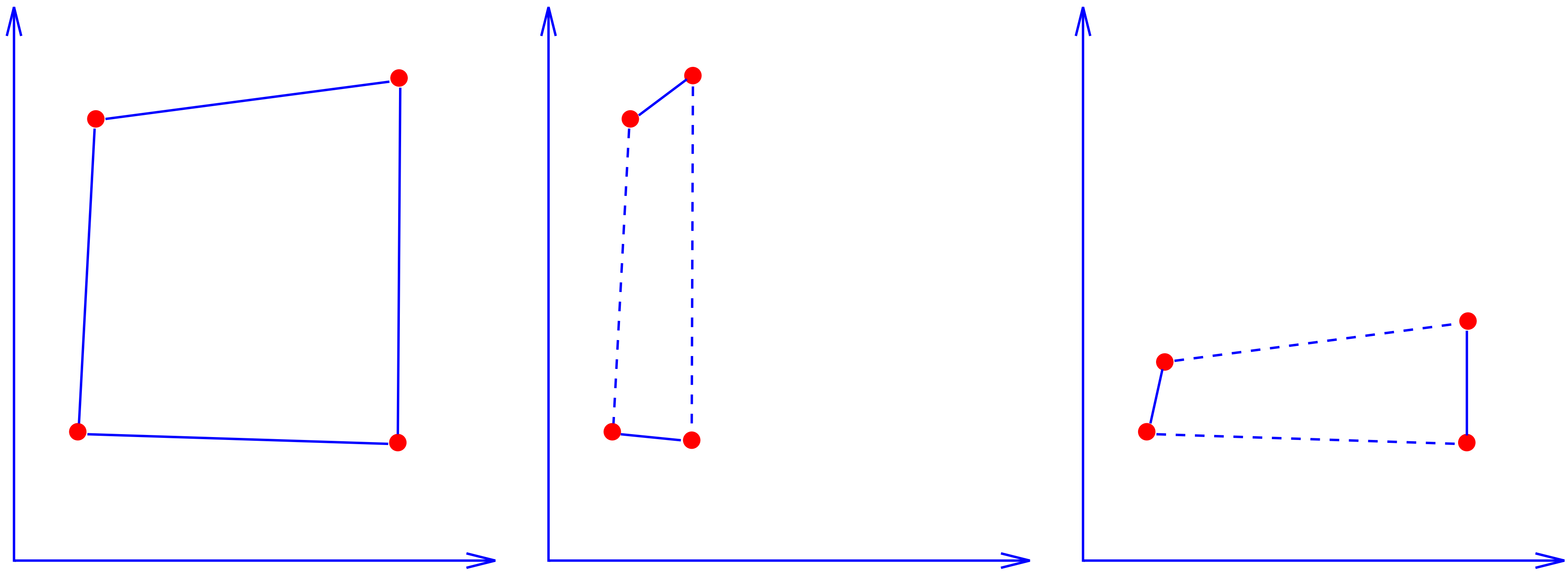
Clustering Issues: Distance Metrics

Distance metrics strongly influence cluster shapes:

- Normalized dot-product: $\frac{\boldsymbol{x}^t \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$
- Euclidean: $\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 = (\boldsymbol{x} - \boldsymbol{\mu}_i)^t (\boldsymbol{x} - \boldsymbol{\mu}_i)$
- Weighted Euclidean: $(\boldsymbol{x} - \boldsymbol{\mu}_i)^t \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{\mu}_i)$ (e.g., $\boldsymbol{W} = \Sigma^{-1}$)
- Minimum distance (chain): $min \quad d(\boldsymbol{x}, \boldsymbol{x}_i), \quad \boldsymbol{x}_i \epsilon C_i$
- Representation specific ...

Clustering Issues: Impact of Scaling

Scaling feature vector dimensions can significantly impact clustering results



Scaling can be used to normalize dimensions so a simple distance metric is a reasonable criterion for similarity

Clustering Issues: Training and Test Data

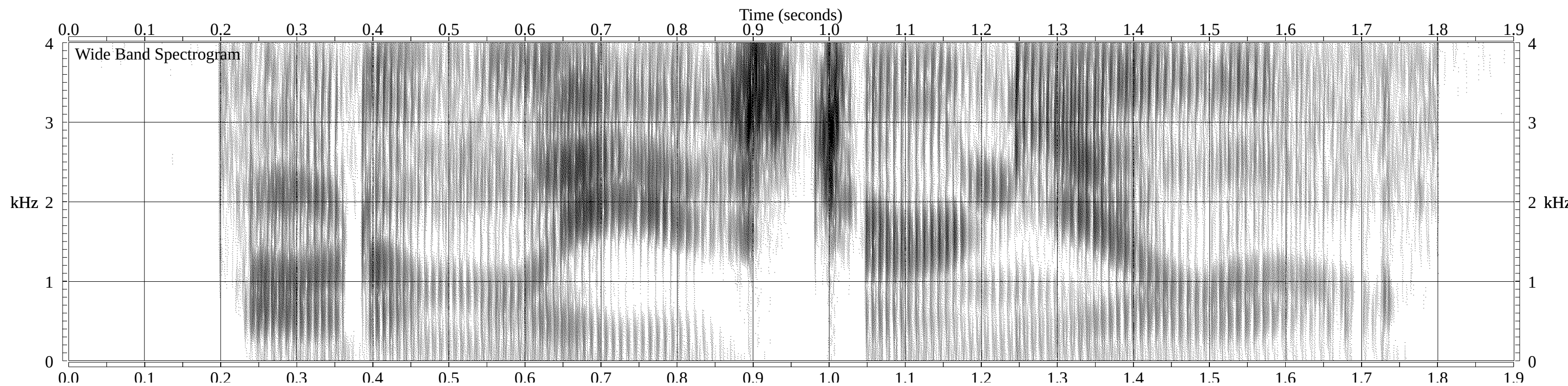
- Training data performance can be arbitrarily good e.g.,

$$\lim_{K \to \infty} \mathcal{D}_K = 0$$

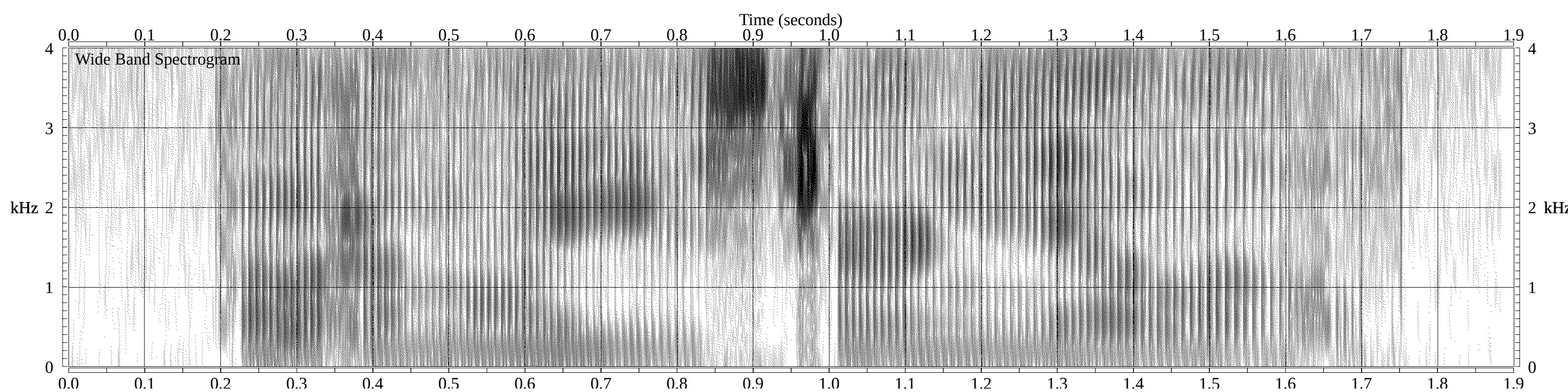
- Independent test data needed to measure performance
  - Performance can be measured by distortion, $\mathcal{D}$, or some more relevant speech recognition metric
  - Robust training will degrade minimally during testing
  - Good training data closely matches test conditions
- Development data are often used for refinements, since through iterative testing they can implicitly become a form of training data

Alternative Evaluation Criterion: LPC VQ Example

Autumn



Autumn LPC



(codebook size = 1024)

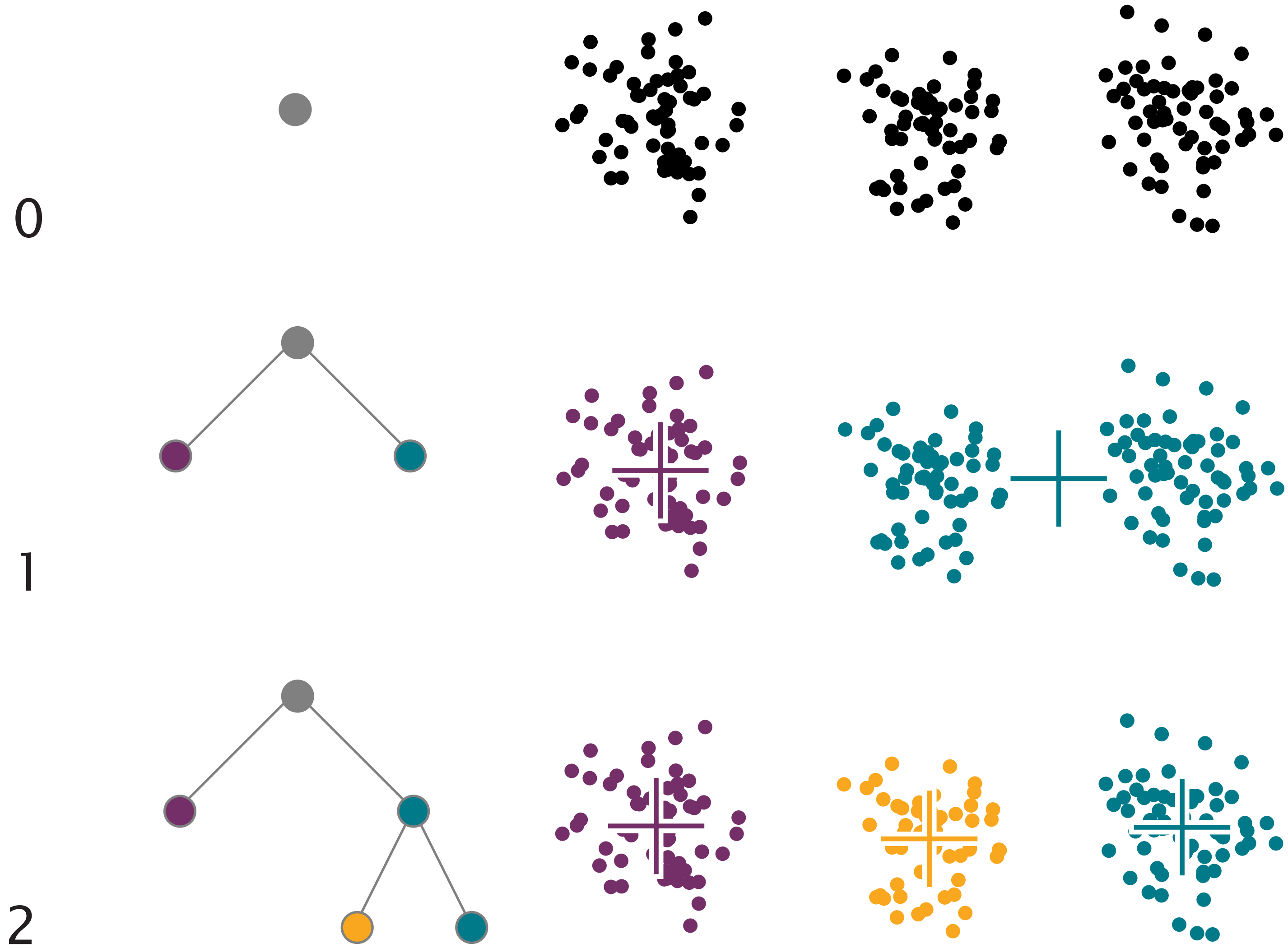
Hierarchical Clustering

- Clusters data into a hierarchical class structure
- Top-down (divisive) or bottom-up (agglomerative)
- Often based on stepwise-optimal, or greedy, formulation
- Hierarchical structure useful for hypothesizing classes
- Used to seed clustering algorithms such as K-means

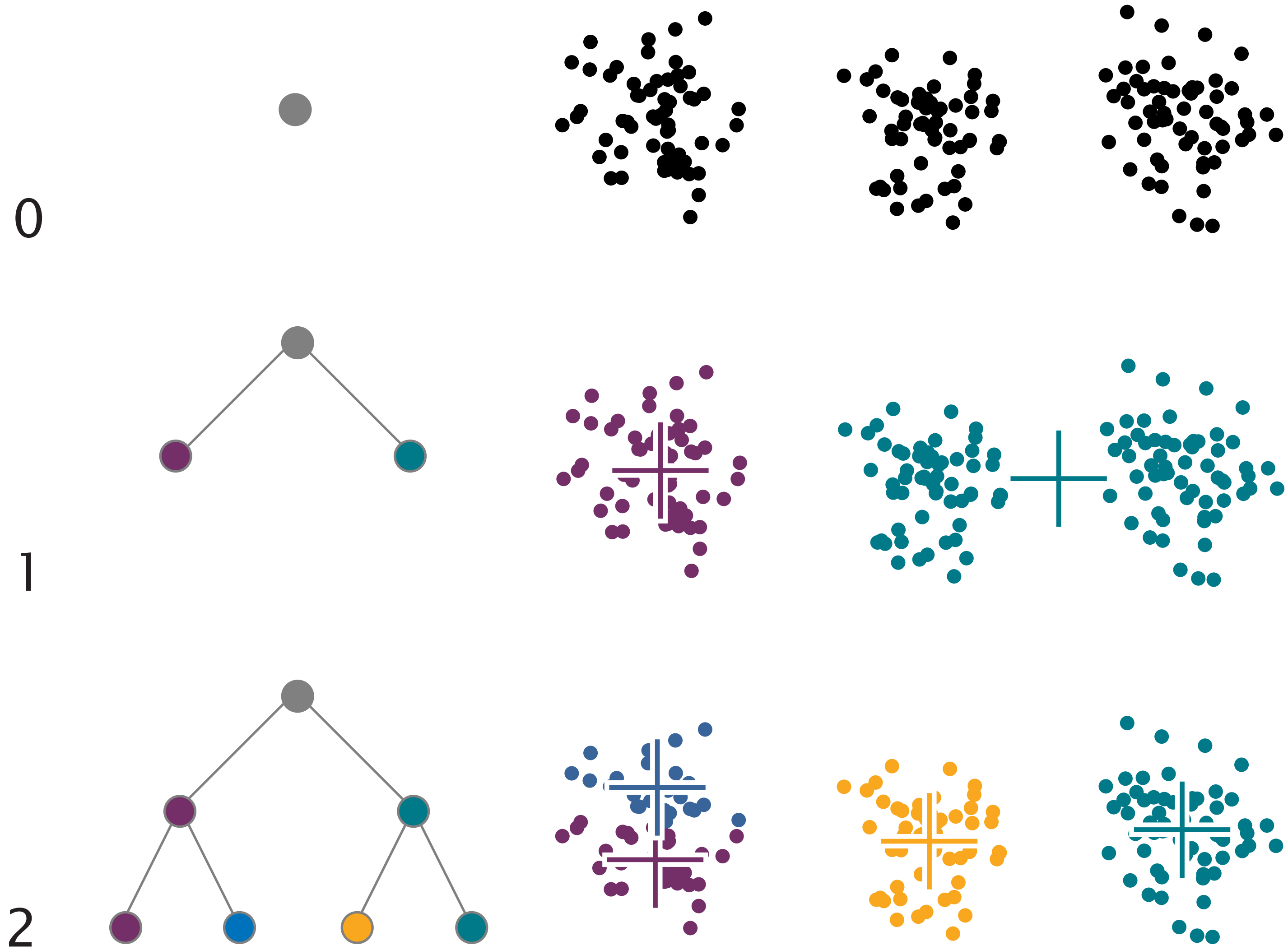
Divisive Clustering

- Creates hierarchy by successively splitting clusters into smaller groups
- On each iteration, one or more of the existing clusters are split apart to form new clusters
- The process repeats until a stopping criterion is met
- Divisive techniques can incorporate pruning and merging heuristics which can improve the final result

Example of Non-Uniform Divisive Clustering



Example of Uniform Divisive Clustering



Divisive Clustering Issues

- Initialization of new clusters
  - Random selection from cluster samples
  - Selection of member samples far from center
  - Perturb dimension of maximum variance
  - Perturb all dimensions slightly
- Uniform or non-uniform tree structures
- Cluster pruning (due to poor expansion)
- Cluster assignment (distance metric)
- Stopping criterion
  - Rate of distortion decrease
  - Cannot increase cluster size

Divisive Clustering Example: Binary VQ

- Often used to create $M = 2^B$ size codebook (B bit codebook, codebook size M)
- Uniform binary divisive clustering used
- On each iteration each cluster is divided in two

$$\boldsymbol{\mu}_i^+ = \boldsymbol{\mu}_i(1 + \epsilon)$$

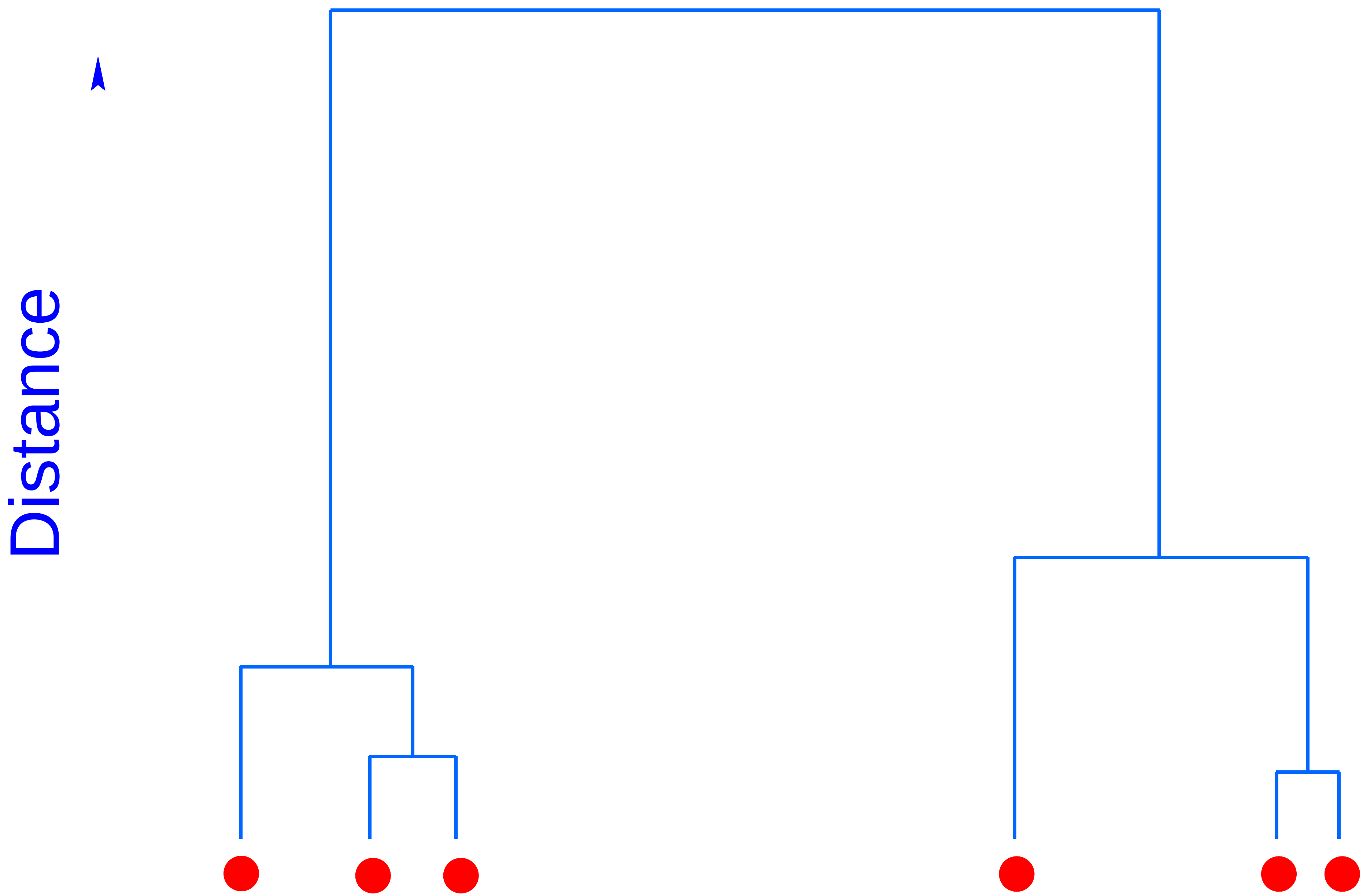
$$\boldsymbol{\mu}_i^- = \boldsymbol{\mu}_i(1 - \epsilon)$$

- K-means used to determine cluster centroids
- Also known as LBG (Linde, Buzo, Gray) algorithm
- A more efficient version does K-means only within each binary split, and retains tree for efficient lookup

Agglomerative Clustering

- Structures N samples or seed clusters into a hierarchy
- On each iteration, the two most similar clusters are merged together to form a new cluster
- After $N - 1$ iterations, the hierarchy is complete
- Structure displayed in the form of a dendrogram
- By keeping track of the similarity score when new clusters are created, the dendrogram can often yield insights into the natural grouping of the data

Dendrogram Example (One Dimension)



Agglomerative Clustering Issues

- Measuring distances between clusters C_i and C_j with respective number of tokens n_i and n_j
  - Average distance: $\frac{1}{n_i n_j} \sum_{i,j} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$
  - Maximum distance (compact): $\max_{i,j} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$
  - Minimum distance (chain): $\min_{i,j} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$
  - Distance between two representative vectors of each cluster such as their means: $d(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$

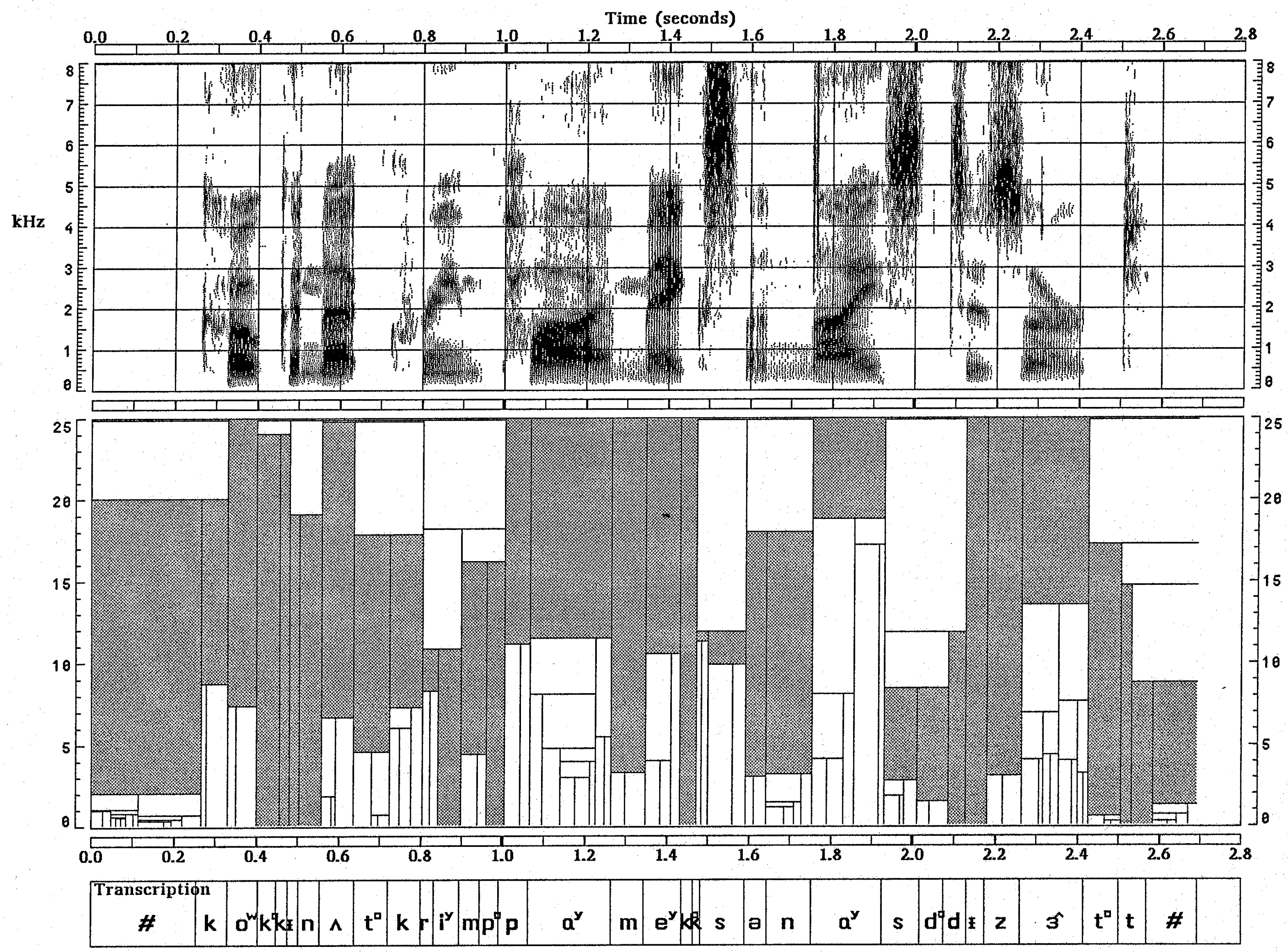
Stepwise-Optimal Clustering

- Common to minimize increase in total distortion on each merging iteration: stepwise-optimal or greedy
- On each iteration, merge the two clusters which produce the smallest increase in distortion
- Distance metric for minimizing distortion, $\mathcal{D}$, is:

$$\sqrt{\frac{n_i n_j}{n_i + n_j}} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$$

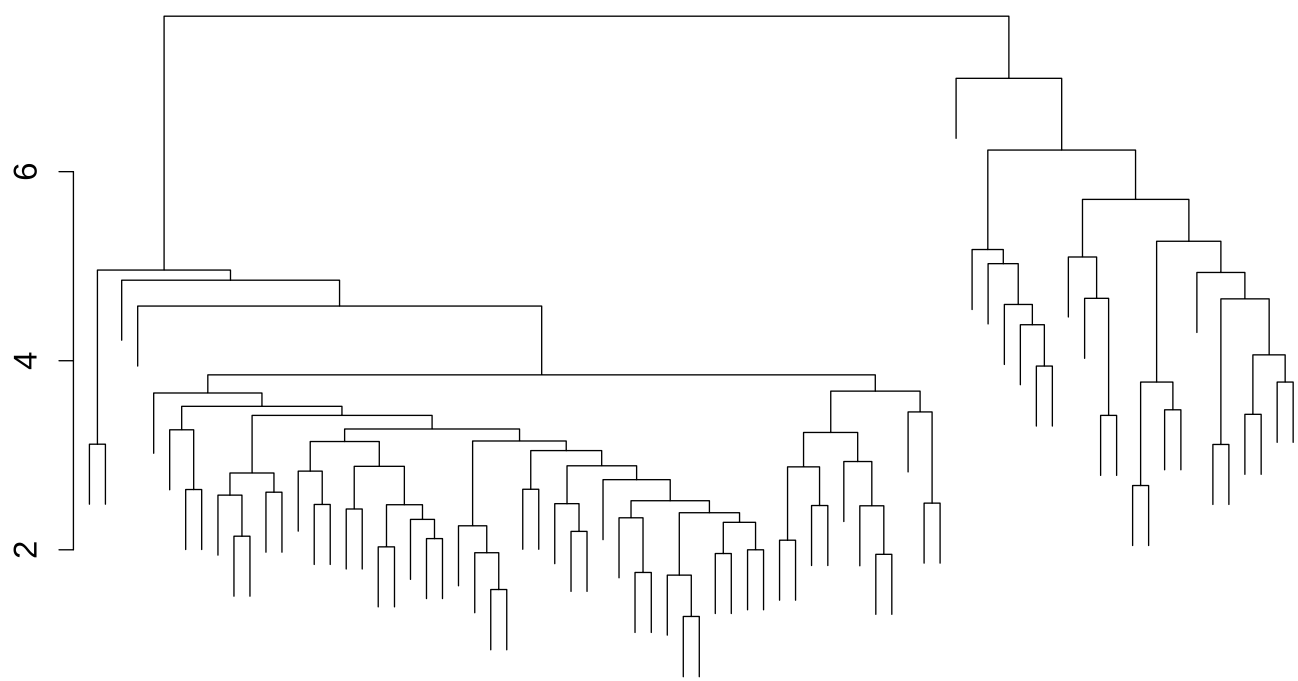
- Tends to combine small clusters with large clusters before merging clusters of similar sizes

Clustering for Segmentation

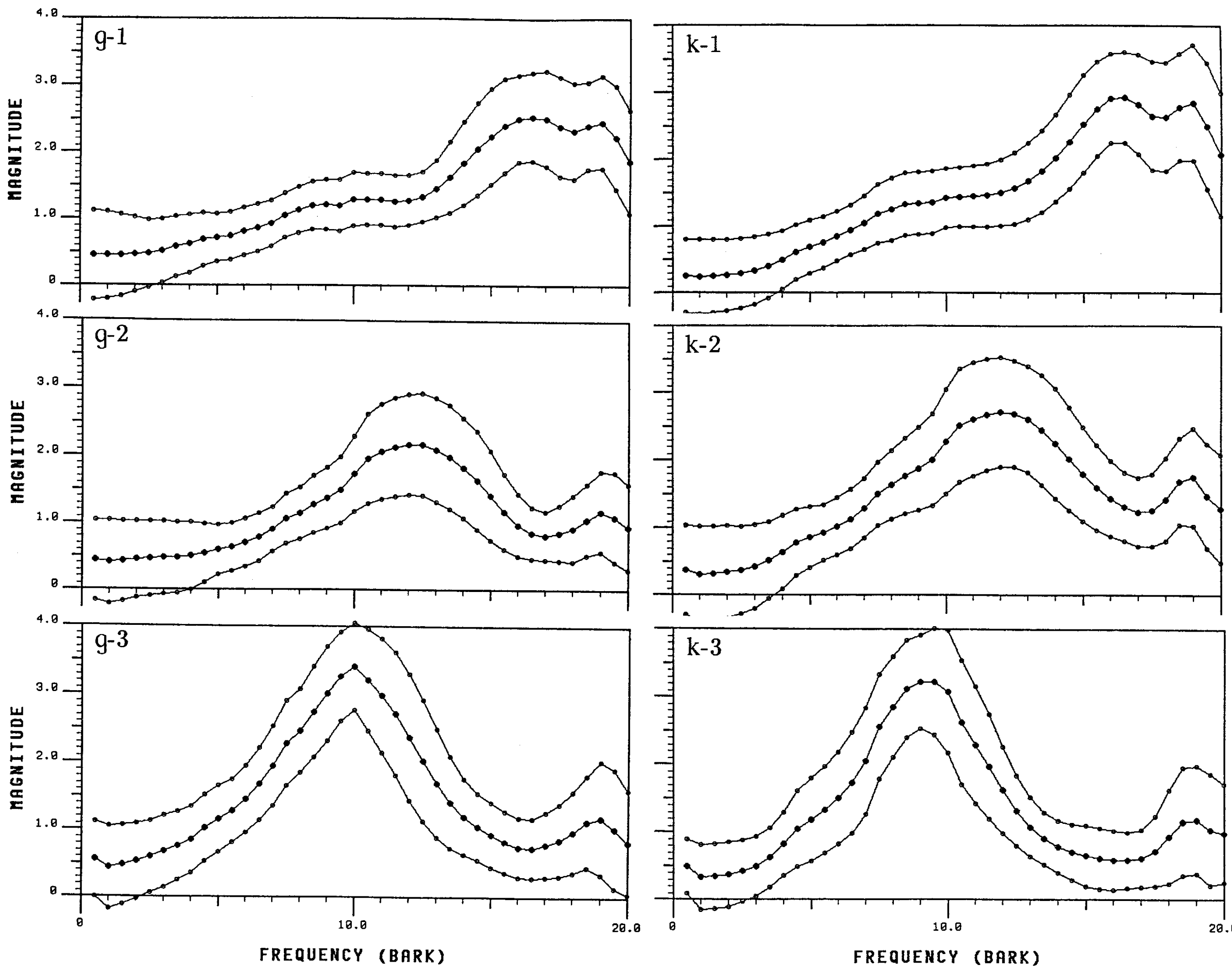


Speaker Clustering

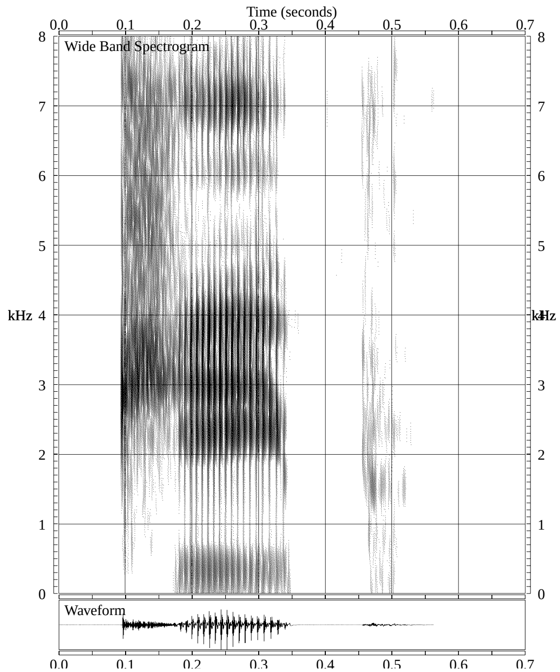
- 23 female and 53 male speakers from TIMIT corpus
- Vector based on F1 and F2 averages for 9 vowels
- Distance $d(C_i, C_j)$ is average of distances between members



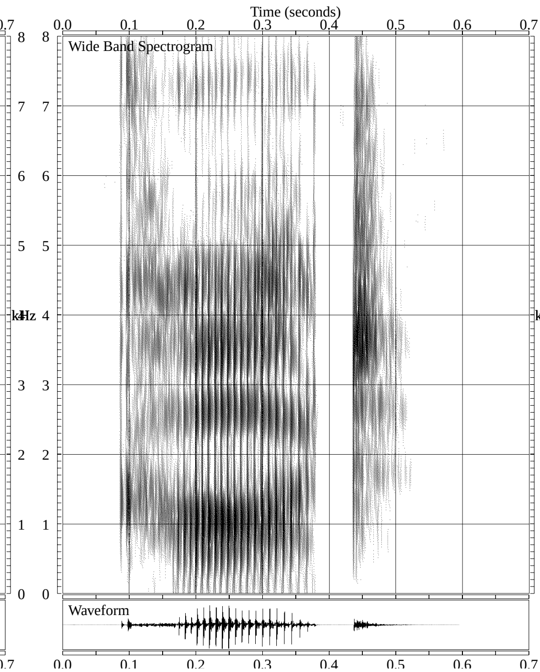
Velar Stop Allophones



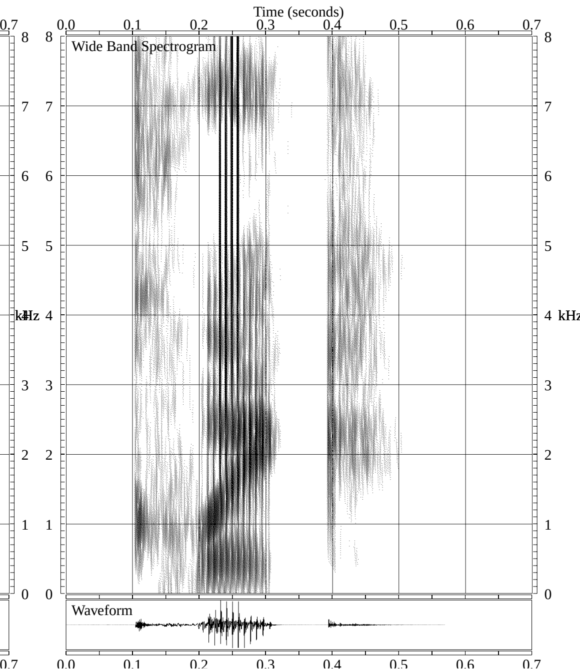
Velar Stop Allophones (con't)



keep



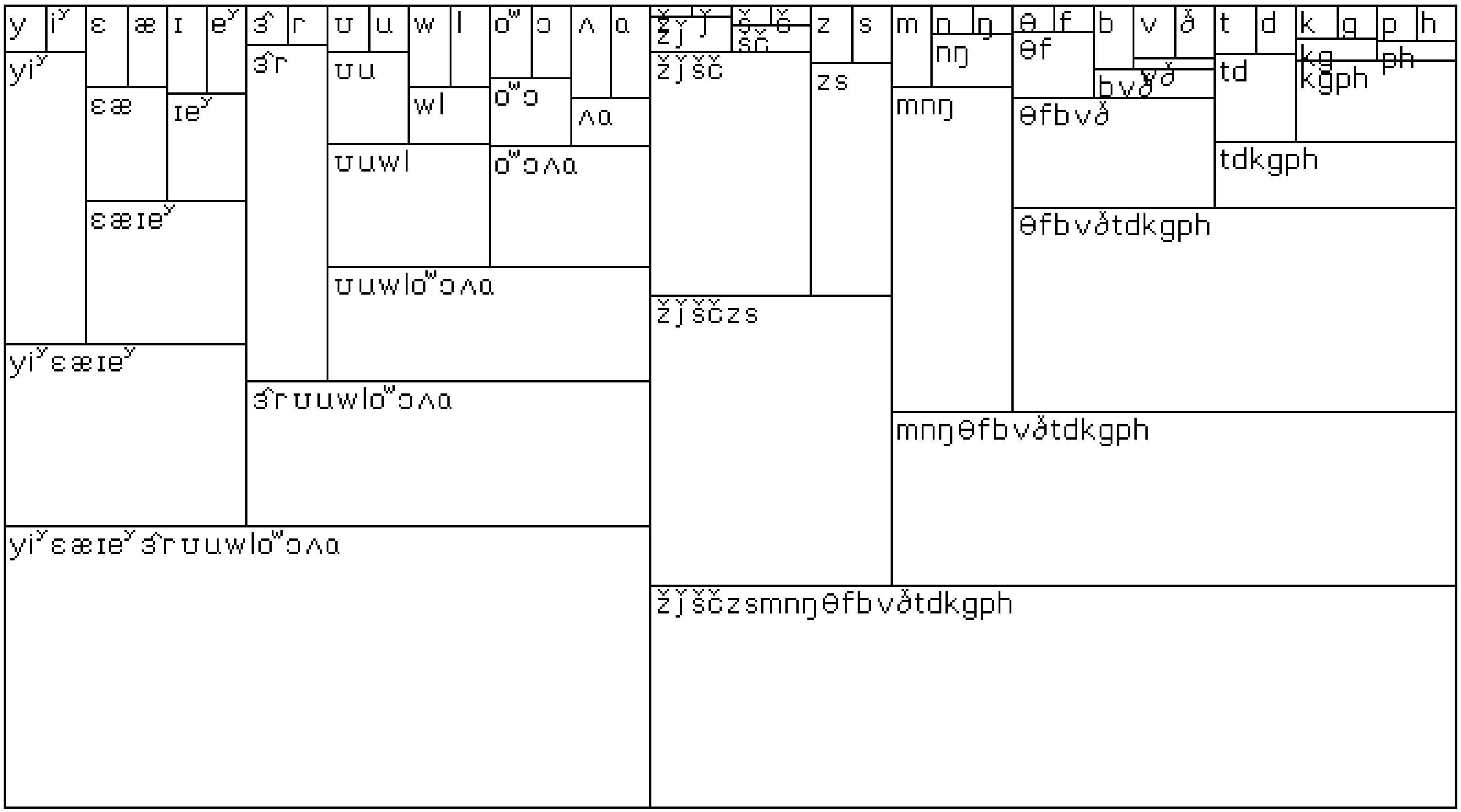
cot



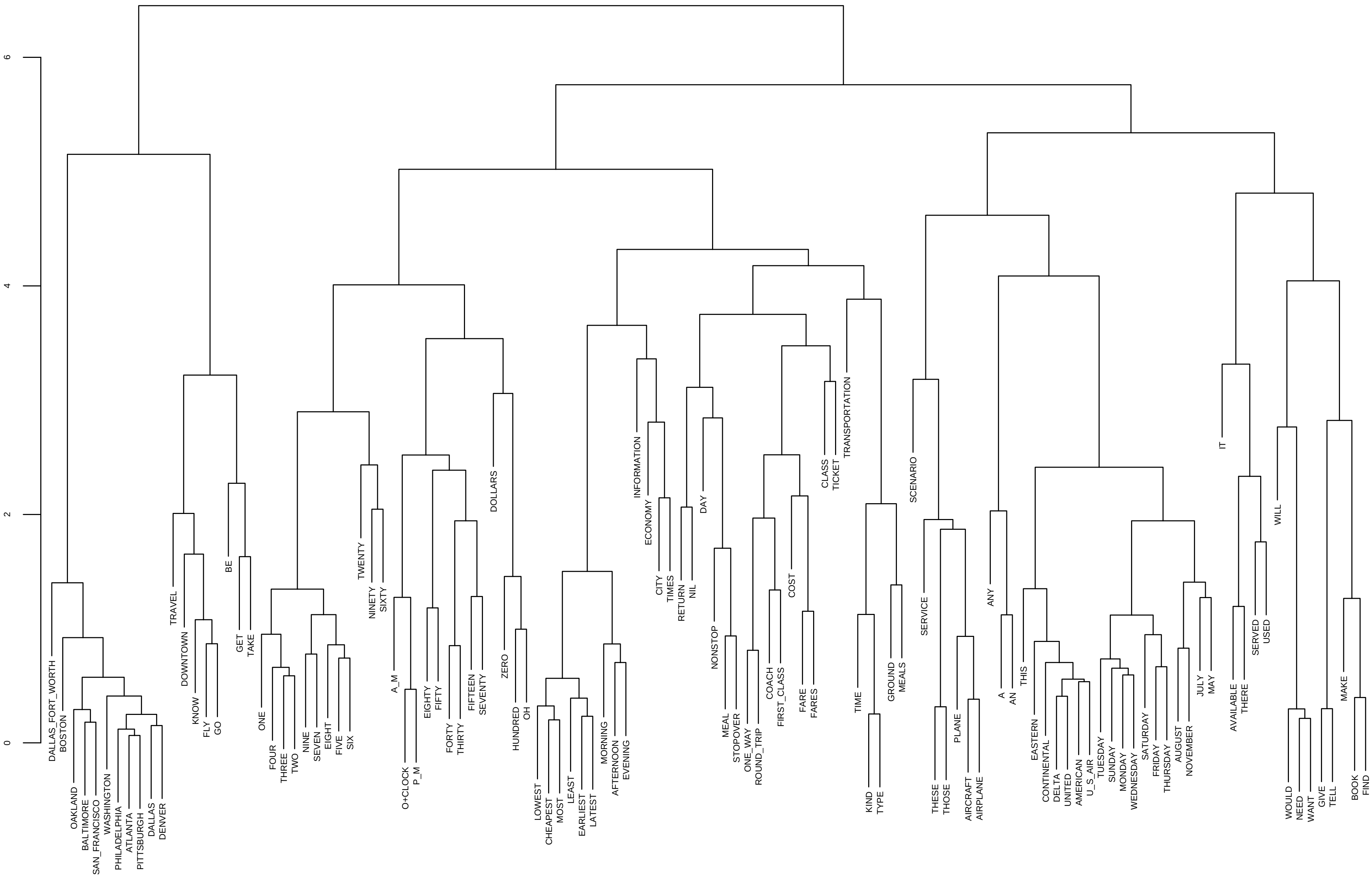
quick

Acoustic-Phonetic Hierarchy

Clustering of phonetic distributions across 12 clusters



Word Clustering



VQ Applications

- Usually used to reduce computation
- Can be used alone for classification
- Used in dynamic time warping (DTW) and discrete hidden Markov models (HMMs)
- Multiple codebooks are used when spaces are statistically independent (product codebooks)
- Matrix codebooks are sometimes used to capture correlation between succesive frames
- Used for semi-parametric density estimation (e.g., semi-continuous mixtures)

References

- Huang, Acero, and Hon, *Spoken Language Processing*, Prentice-Hall, 2001.
- Duda, Hart and Stork, *Pattern Classification*, John Wiley & Sons, 2001.
- A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Press, 1992.
- R. Gray, Vector Quantization, *IEEE ASSP Magazine*, 1(2), 1984.
- B. Juang, D. Wang, A. Gray, Distortion Performance of Vector Quantization for LPC Voice Coding, *IEEE Trans ASSP*, 30(2), 1982.
- J. Makhoul, S. Roucos, H. Gish, Vector Quantization in Speech Coding, *Proc. IEEE*, 73(11), 1985.
- L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.