Speech Recognition for English E-Set

- E-Set: {b, c, d, e, g, p, t, v, z}, total nine English letters with the same ending vowel “E”.
- Speech Recognition: conducted in a multi-speaker, isolated-word mode.
- Training Set: 900 tokens from 100 speakers
- Testing Set: another 900 tokens from the same 100 speakers
- Speech Signal:
  - Sampled at 6.67 kHz,
  - Processed by using an eighth order LPC analysis, with a 45 ms window and a 15 ms shift.
  - A 24-component feature vector, consisting of 12 LPC-derived liftered cepstral coefficients and 12 delta cepstral coefficients.

Speech Recognition Baseline Model:

- Each letter is modeled by a 5-state left-to-right HMM
- Five Gaussian mixture components per state are used for characterizing each state observation density.
- The HMM parameters were estimated using the segmental k-means algorithm
- Recognition Rates: 61.7% for the testing set, and 80.2% for the training set.

Discrimination-based Approaches [Su and Lee 94]

- Input observation vector X: $X^t = \left[S_1^1, S_2^1, S_3^1, \cdots S_3^9, S_4^9, S_5^9 \right]$
  - where S^i_j is either the averaged or the accumulated state log-likelihood (score) for state *j* or word *I.* There are 45 elements in this *score vector*.
- Subspace Projection:
  - Maximin Algorithm is proposed for feature selection:
    - Divergence $D_{i,j}(k)$ between the class i and the class j, for 1 <= i, j <= 9; i < j, 1 <= k <= 45.

$$D_{ij}(k) = \int_{Y_k} (P_i(Y_k) - P_j(Y_k)) \cdot \ln \left(\frac{P_i(Y_k)}{P_j(Y_k)} \right) dY_k$$

    - Find $D_{min}(k) = \min_{i,j} D_{i,j}(k)$ for 1 <= i, j <= 9; i < j, 1 <= k <= 45.
    - Sort $D_{min}(k)$ in descending order, then the sequence of subspaces is obtained

Discrimination-based Approaches (cont.):

- Subspace Projection (cont.):
  - Result:
    - Table I: The 45 Features Indices Listed in the Order of Descending Divergence Values (Left to Right First, then Top to Bottom)
    - Table II: Corresponding Divergence Values Listed in Descending Order
    - Figure 1: Recognition rate as a function of dimensionality for a subspace-based recognizer
    - Figure 2: Recognition rate as a function of iteration for a robust subspace-based recognizer

Discrimination-based Approaches (cont.):

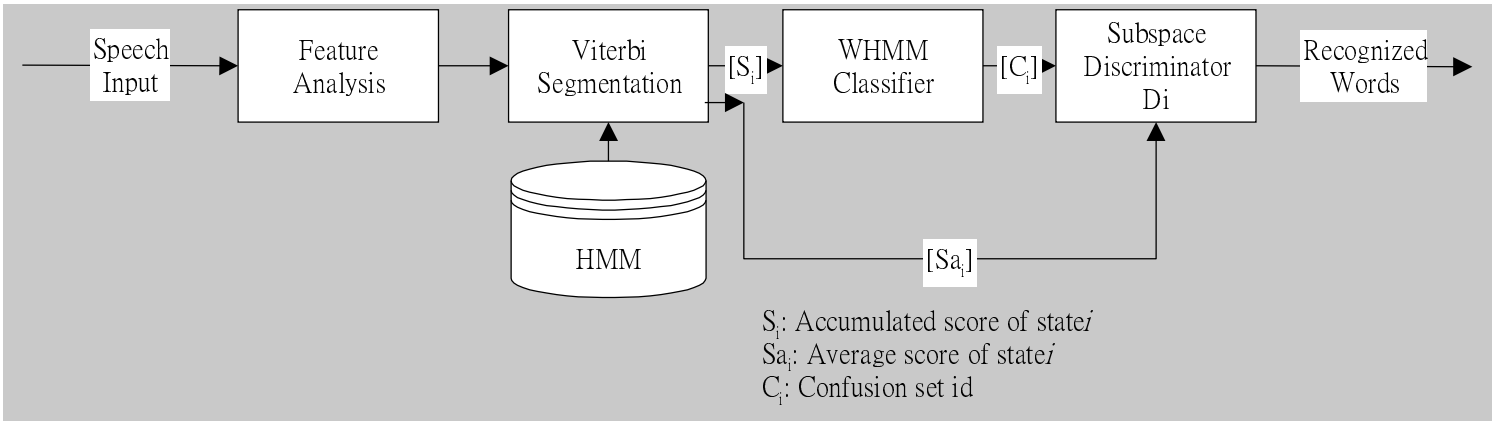
- Weighted HMM Algorithm:
  - A linear discriminator is used: $g\left(X, W_i\right) = W_i^t \cdot X$

$$\hat{C} = \arg\max_i \left(W_i^t \cdot X\right)$$

  - Result:
    - Figure 3: Recognition rate as a function of iteration number for a WHMM-based recognizer
    - Figure 4: Recognition rate as a function of iteration number for a robust WHMM recognizer
    - Figure 5: Recognition rate as a function of margin ratio for a robust WHMM recognizer

Discrimination-based Approaches (cont.):

- A Two-Stage Discriminator:
  - Block Diagram: A block diagram of the proposed two-stage recognition system



  - Result:

    - Table III: Performance of Each Confusion Set in Each Stage

Discrimination-based Approaches (cont.):

TABLE I

THE 45 FEATURE INDICES LISTED IN THE ORDER OF DESCENDING DIVERGENCE VALUES (LEFT TO RIGHT FIRST, THEN TOP TO BOTTOM)

1 (B_1)	6 (C_1)	31 (T_1)	41 (Z_1)	26 (P_1)
21 (G_1)	17 (E_2)	37 (V_2)	36 (V_1)	16 (E_1)
22 (G_2)	18 (E_3)	11 (D_1)	38 (V_3)	2 (B_2)
12 (D_2)	10 (C_5)	4 (B_4)	28 (P_3)	43 (Z_3)
30 (P_5)	20 (E_5)	34 (T_4)	27 (P_2)	7 (C_2)
32 (T_2)	19 (E_4)	14 (D_4)	33 (T_3)	45 (Z_5)
13 (D_3)	9 (C_4)	40 (V_5)	39 (V_4)	42 (Z_2)
25 (G_5)	15 (D_5)	24 (G_4)	8 (C_3)	29 (P_4)
5 (B_5)	3 (P_3)	23 (G_3)	35 (T_5)	44 (Z_4)

Discrimination-based Approaches (cont.):

TABLE II

THE 45 CORRESPONDING DIVERGENCE VALUES LISTED IN DESCENDING ORDER

0.099410	0.066085	0.042937	0.038249	0.037496
0.028690	0.020272	0.019982	0.017487	0.011872
0.009741	0.005572	0.005305	0.004210	0.004037
0.004005	0.003803	0.003728	0.003656	0.003621
0.003214	0.003200	0.003085	0.002916	0.002690
0.002147	0.002064	0.001937	0.001900	0.001661
0.001651	0.001637	0.001496	0.001221	0.000969
0.000969	0.000596	0.000562	0.000384	0.000381
0.000229	0.000194	0.000187	0.000076	0.000042

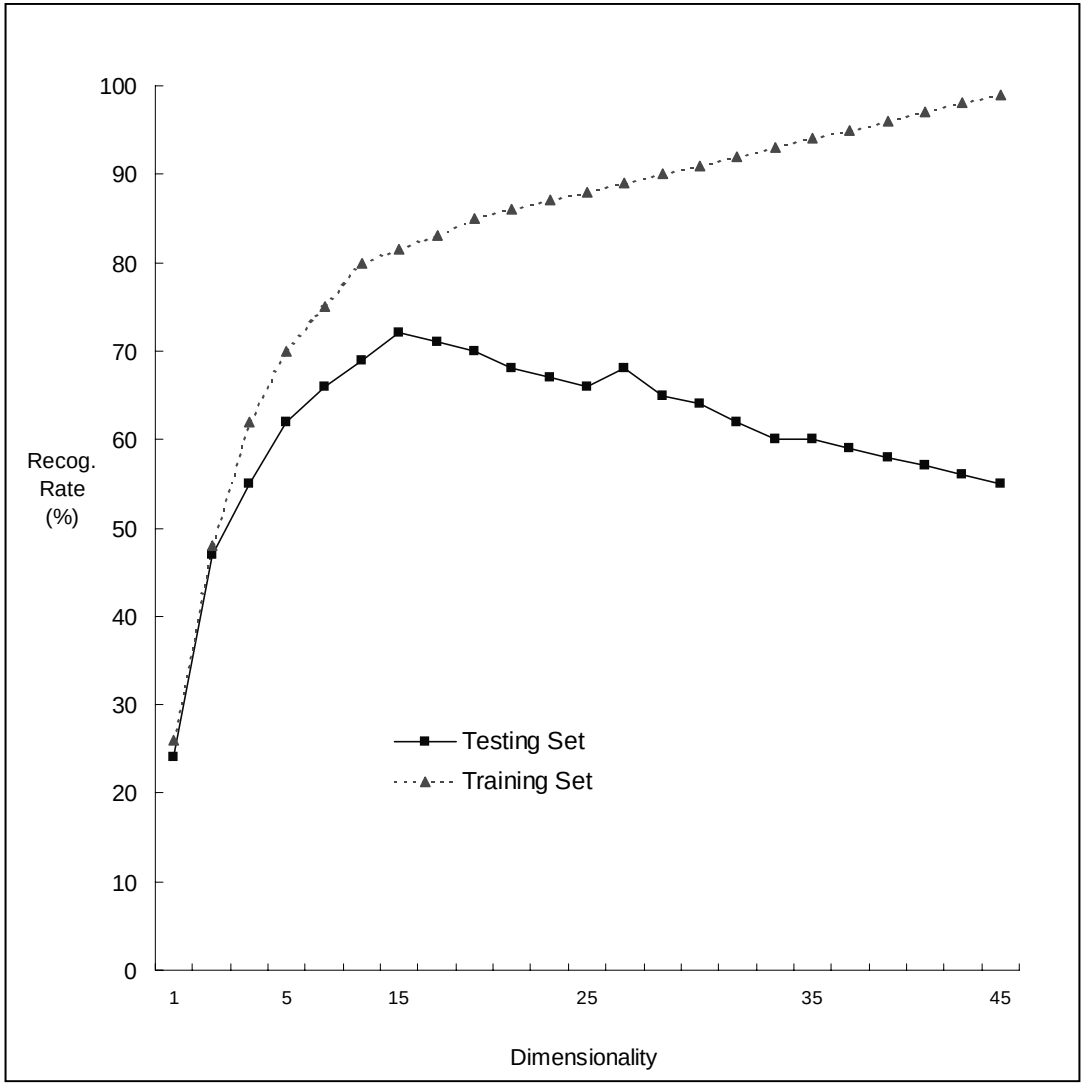
Discrimination-based Approaches (cont.):

TABLE III

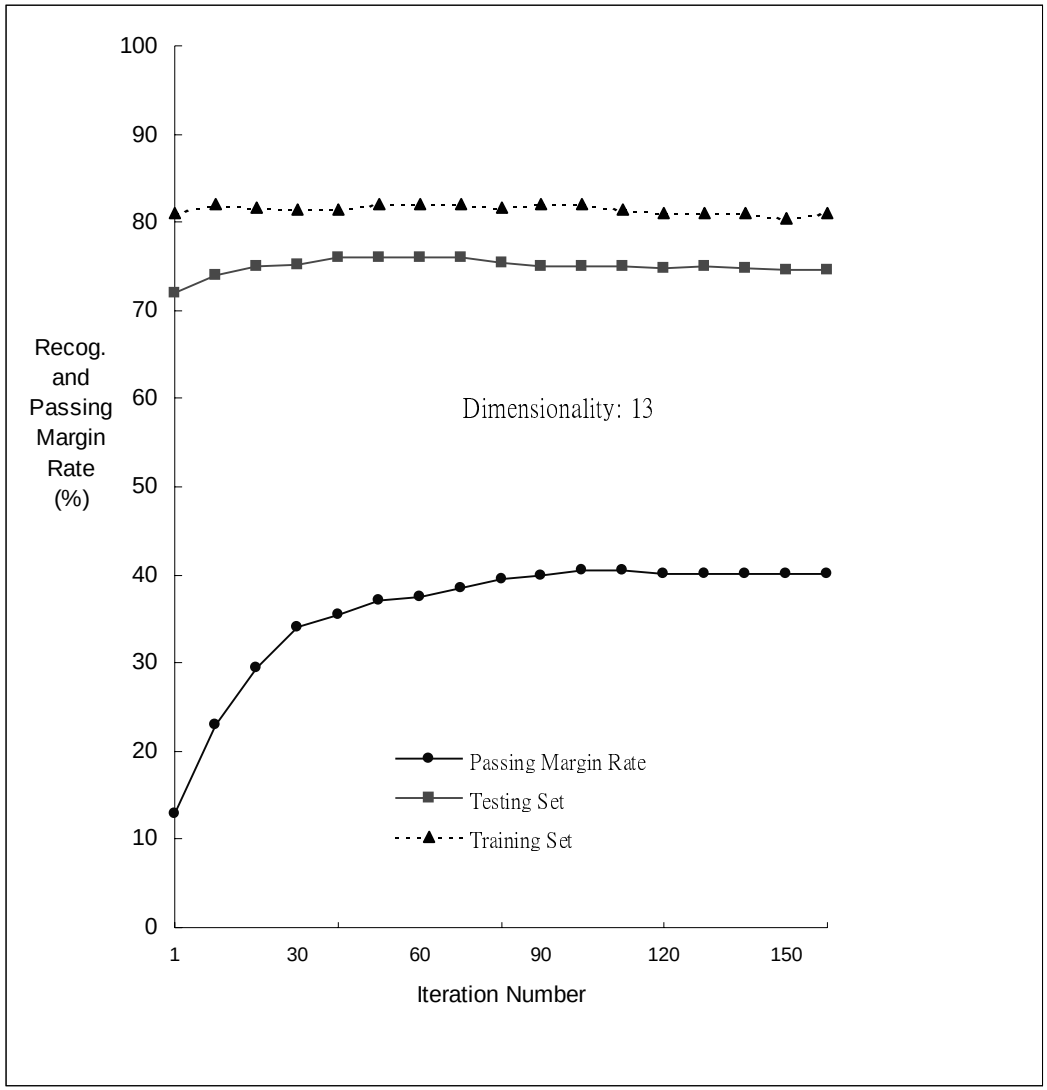
PERFORMANCE OF EACH CONFUSION SET IN EACH STAGE

	First Stage (%)	Second Stage (%)
B	68.9	74.7
C	88.9	89.6
D	64.3	73.9
E	87.6	89.7
G	79.6	84.5
P	72.0	73.1
T	68.0	75.7
V	73.2	75.6
Z	73.3	78.1
Average	74.9	79.4

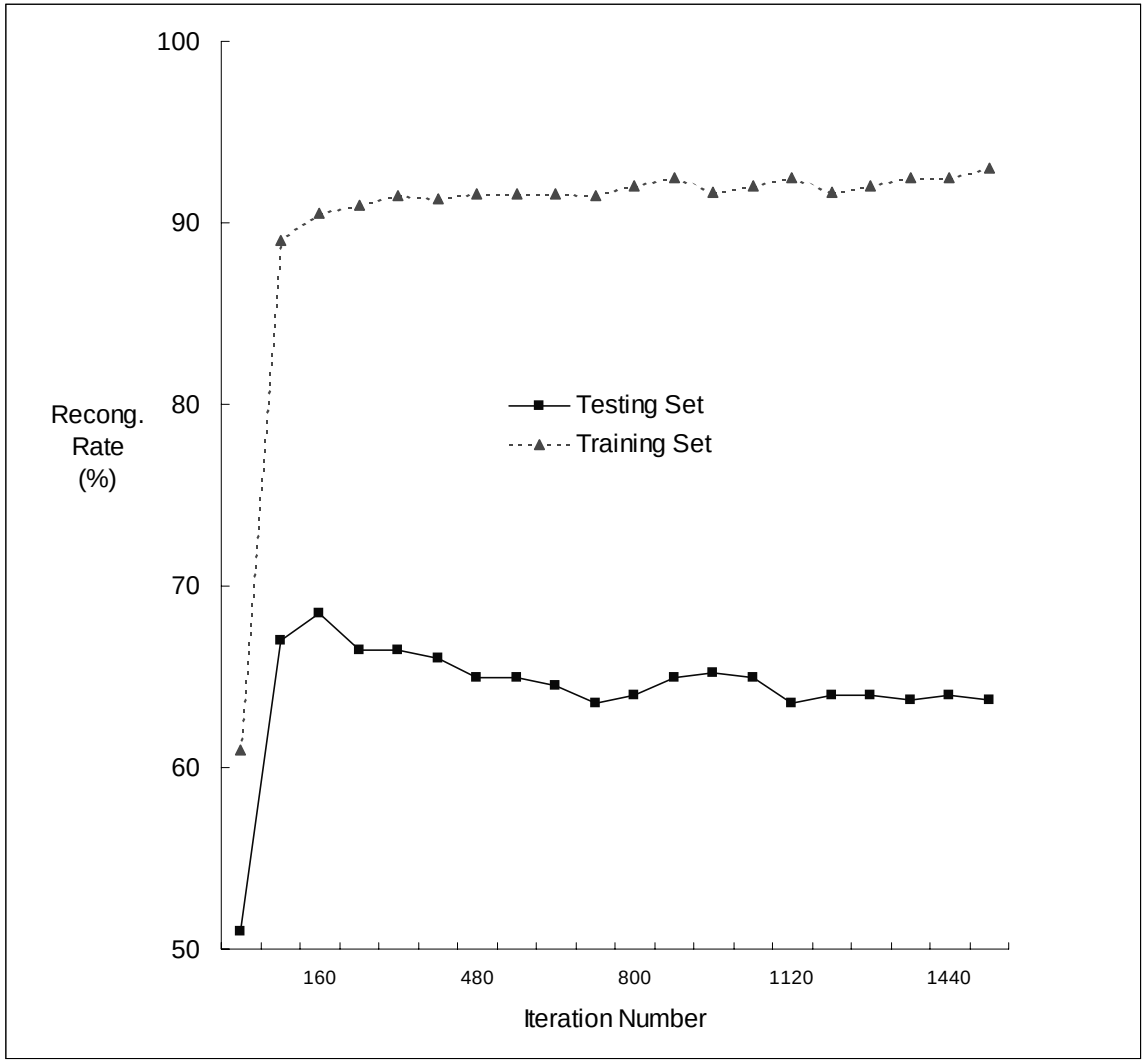
Subspace-based Recognizer: Recognition Rate vs. Dimensionality



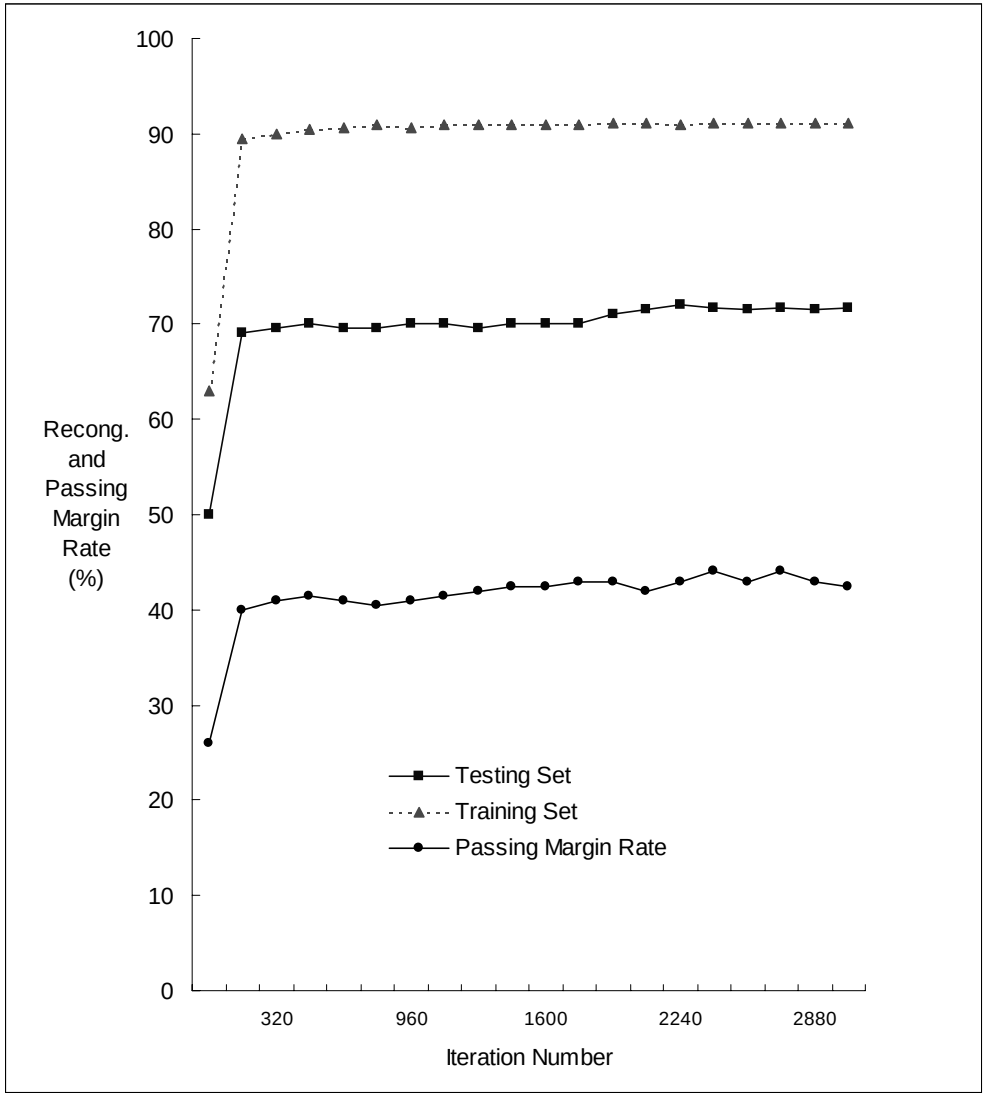
Robust Subspace-based Recognizer: Recognition Rate vs. Iteration



WHMM-based Recognizer: Recognition Rate vs. Iteration Number



Robust WHMM Recognizer: Recognition Rate vs. Iteration Number



Robust WHMM Recognizer: Recognition Rate vs. Margin Ratio