Rapid Speaker Adaptation in Eigenvoice Space

by R. Kuhn, J-C. Junqua, P. Nguyen, and N. Niedzielski
(IEEE Trans. SAP, Nov. 2000)

Heeyoul (Henry) Choi
Dept. of Computer Science
Texas A&M University
hchoi@cs.tamu.edu

Outline

- Introduction
  - Speaker Adaptation
- Eigenvoice
- Comparison with others
  - MAP, MLLR, EMAP, RMP, CAT, RSW…
- Experiments
- Future work
- Summary

Introduction

- Speaker-dependent (SD) system
- Speaker-independent (SI) system

- Speaker Adaptation
  - Finding SD system for a new speaker with small data.

- This paper is about making the adaptation faster based on eigenvoice approach.

Speaker Adaptation

- Model-based algorithms
  - Adapt to a new speaker by modifying the parameters of the system's speaker model.
- Maximum a posteriori (MAP), Maximum likelihood linear regression (MLLR).
  - Require significant amounts of adaptation data from the new speaker.

Speaker Adaptation

- Speaker space algorithm
  - Constrain the adapted model to be a linear combination of a small number of basis vectors from the reference speakers.
  - Faster and robust
    - Related to speaker clustering in fact that they reduce the parameter dimension to search.
    - Resemble extended MAP (EMAP) in fact that they use a priori information from reference speakers.
    - Actually, prior information is used to reduce the parameter space.
- Eigenvoice is one of these algorithm.

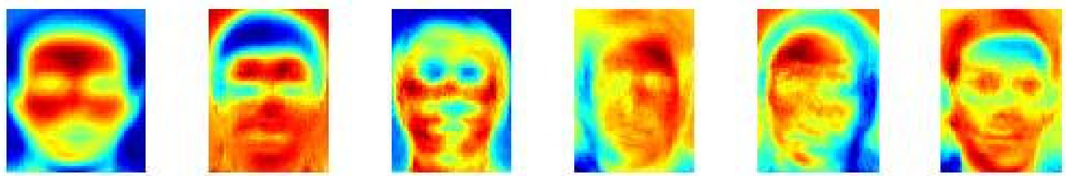
Speaker Adaptation

- Eigenvoice
  - Finds basis vectors that are orthogonal to each other
  - Efficient in the sense of variation.
  - Has all property of principal component analysis (PCA)

  - PCA is applied to the parameter space.

Eigenface

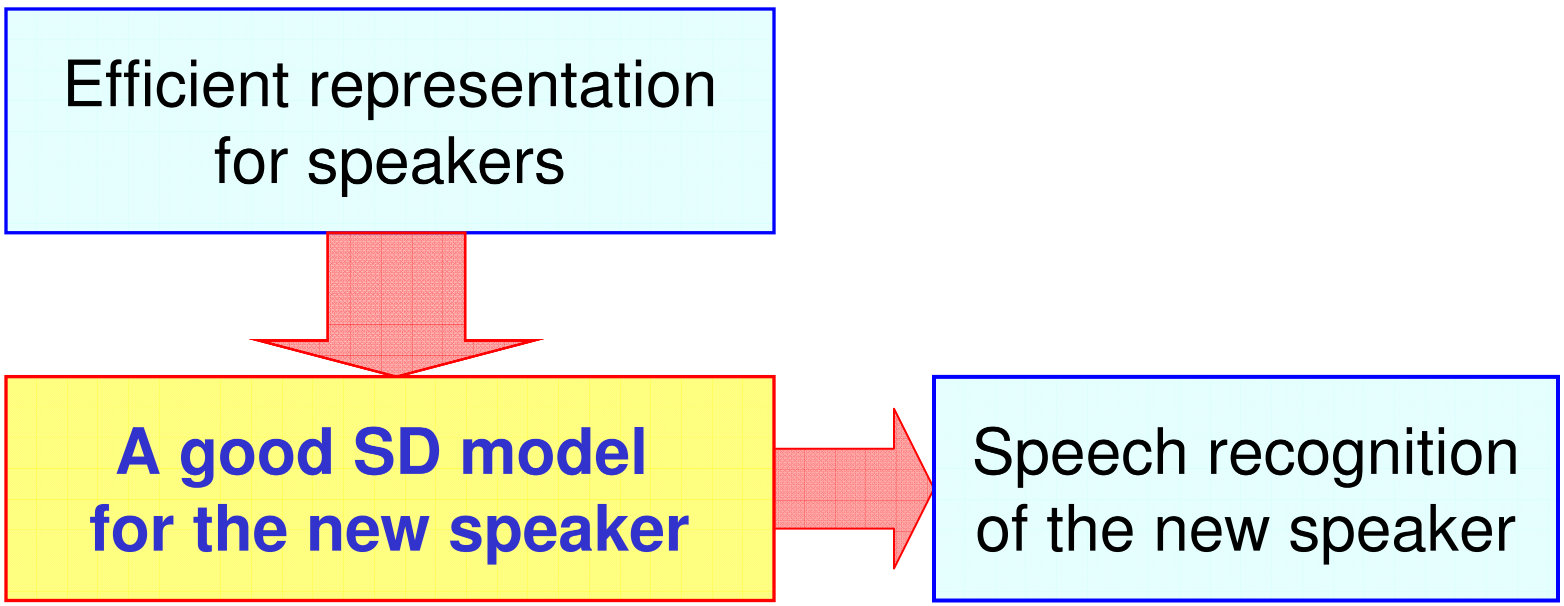
- Eigenvoice is an analogy to eigenface in face images.
  - Face is a weighted sum of eigenfaces, which are eigenvectors of face images.



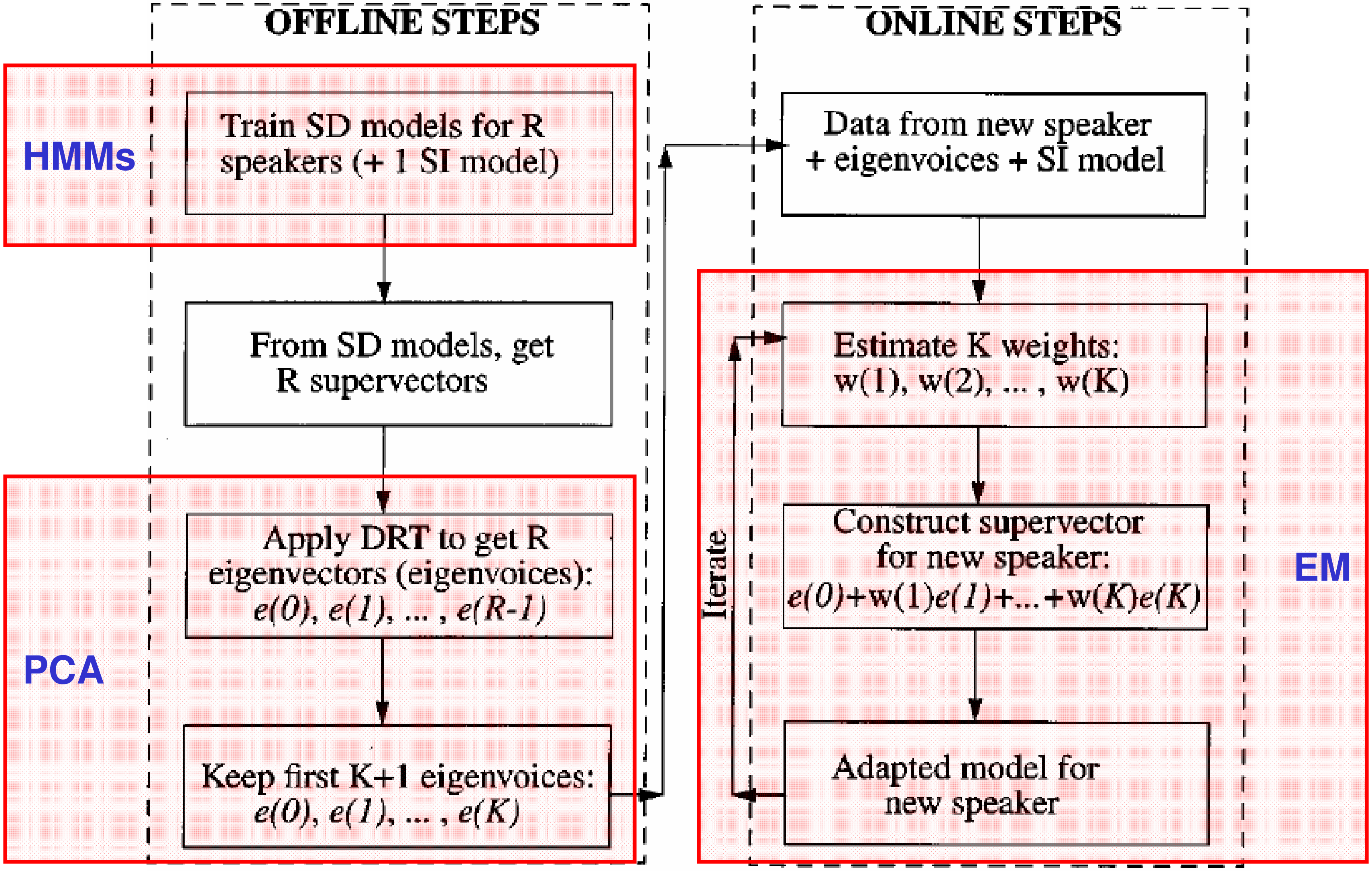
- PCA
  - Ordered by eigenvalues.
  - Guarantee the minimized mean-square error.

Eigenvoice

- Face recognition => speaker recognition? How?
- Speaker recognition vs. speech recognition



Eigenvoice



OFFLINE STEPS
ONLINE STEPS
HMMs
Train SD models for R speakers (+ 1 SI model)
From SD models, get R supervectors
PCA
Apply DRT to get R eigenvectors (eigenvoices): e(0), e(1), ... , e(R-1)
Keep first K+1 eigenvoices: e(0), e(1), ... , e(K)
Data from new speaker + eigenvoices + SI model
Estimate K weights: w(1), w(2), ... , w(K)
Construct supervector for new speaker: e(0)+w(1)e(1)+...+w(K)e(K)
Adapted model for new speaker
Iterate
EM

Eigenvoice (HMM)

- Supervector : Model parameters.
  - The means of HMM output Gaussians.
  - Not voice. (Actually, eigen_model_parameter)
- Instead of PCA,
  - independent component analysis (ICA) (Factorialvoice as in factorialface)
  - linear discriminant analysis (LDA) (generalized eigenvoice)
- They used correlation matrix instead of covariance matrix.

Eigenvoice (HMM)

- Hidden states of every SD model are from SI model.
  - Insufficient data for SD but enough for SI.
  - (SI + small data) is enough to make SD model.

- Hidden states means kind of Speaker invariant features.
  - Each speaker's characters are from the mixture of Gaussian for each state after adaptation.
  - This makes sense to build new speakers model with the same states and transition probabilities as SI model.

Eigenvoice (PCA)

- Now, the adaptation procedure is redefined as finding K parameters.

- Computationally cheap.
  - Only the weights of eigenvoices instead of all parameters.
- Requires only a small amount of adaptation data
  - Think about the ‘curse of dimensionality’
- And robust against noise
  - The discarded eigenvectors corresponding to small eigenvalues might be about ‘noise’.

Eigenvoice (EM-Initialization)

- Initialization of weights and the others (variances and transition probabilities) are from SI model

- The parameters for the new speaker is

$$P = e(0) + w(1) * e1 + \cdots + w(K) * e(K).$$

  - The problem is to estimate the weight *w(j)* from data.

- Maximum likelihood eigen-decomposition (MLED)
  - Gaussian mean adaptation in a continuous density hidden Markov model (CDHMM).

Eigenvoice (EM-MLED)

- Likelihood (lamda ={means of Gaussian})

$$P(O \mid \lambda) = \sum_{m,s} P(O, m, s \mid \lambda) = \sum_{m,s} P(m, s \mid \lambda) P(O \mid m, s, \lambda)$$

- Auxiliary function

$$Q(\lambda, \hat{\lambda}) = \sum_{m,s} P(O, m, s \mid \lambda) \log P(O, m, s \mid \hat{\lambda})$$

$$P(O, m, s \mid \hat{\lambda}) = \prod_t P(o_t, m, s \mid \hat{\lambda}) = \prod_t P(o_t \mid m, s, \hat{\lambda}) P(m, s \mid \hat{\lambda})$$

  - $\gamma_m^{(s)}(t) = P(m, s \mid \lambda, \mathbf{o}_t)$ (s-m occupation prob.)
  - $P(o_t \mid m, s, \hat{\lambda})$ is a Gaussian distribution.

Eigenvoice (EM-MLED)

- Finally,

$$Q(\lambda, \hat{\lambda}) = -\frac{1}{2} P(O|\lambda) \times \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) f(\mathbf{o}_t, s, m)$$

where

$$f(\mathbf{o}_t, s, m) = [-n \log(2\pi) - \log |C_m^{(s)}| + h(\mathbf{o}_t, s, m)]$$

and

$$h(\mathbf{o}_t, s, m) = \left(\mu_m^{(s)} - \mathbf{o}_t\right)^T C_m^{(s)-1} \left(\mu_m^{(s)} - \mathbf{o}_t\right)$$

Eigenvoice (EM-MLED)

- In the Gaussians, the mean estimates are

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_1^{(1)} \\ \hat{\mu}_2^{(1)} \\ \vdots \\ \hat{\mu}_m^{(s)} \\ \vdots \end{bmatrix} = \sum_{j=1}^{K} w(j)e(j)$$

$$\mu_m^{(s)} = \sum_{j=1}^{K} w(j)e_m^{(s)}(j)$$

  - Pa m the D-dimensional means into the K-dimensional weights (K << D).

Eigenvoice (EM-MLED)

To maximize $Q(\lambda, \hat{\lambda})$, set $(\partial Q / \partial w(j)) = 0$, $j = 1 \cdots K$; assuming the eigenvalues are independent, $(\partial w(i) / \partial w(j)) = 0$, $i \neq j$. One obtains for $j = 1 \cdots K$

$$\sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \left(e_m^{(s)}(j) \right)^T C_m^{(s)-1} \mathbf{o}_t$$
$$= \sum_s \sum_m \sum_t \gamma_m^{(s)}(t)$$
$$\times \left\{ \sum_{k=1}^{K} w(k) \left(e_m^{(s)}(k) \right)^T C_m^{(s)-1} e_m^{(s)}(j) \right\}$$

Comparison (MAP and MLLR)

- Model based algorithm.
- Maximum a Posteriori (MAP)
  - Uses the prior information of parameters in Bayes rule.
  - Updates only the parameters of Gaussians that have observations o_t
  - The number of parameters are large.
- Maximum Likelihood Linear Regression (MLLR)
  - Update all the parameters which is formulated by the linear regression.
  - Much less constrained by prior knowledge. A little by SI model.
  - The number of parameters are small. (the transformation matrix)

- Eigenvoice
  - puts a heavy emphasis on prior knowledge by eigenvectors
  - updates all the parameters by the weights. (Eigenvector is D-dimension)
  - The number of parameters are smaller than MLLR.

Comparison (RMP)

- Extended Maximum a Posteriori (EMAP)
  - Faster convergence by the correlations between observation.
    - Update all the correlated parameters with one observation as much as they are related.
  - Regression-based model prediction (RMP)
    - EMAP to CDHMM
    - Still faster than MAP.
    - Better performance than MAP.
    - Kind of a mixture of MLLR and MAP

Comparison (Clustering)

- Hard speaker clustering
  - Clusters of reference speakers – SI models (cf. codewords)
  - When the new speaker data is available, choose one model.
  - And then MLLR can be used.

- Soft speaker clustering
  - The new speaker's model is a linear combination of reference speaker's models.
  - Clustering + MLLR
    - Clustering works as prior to MLLR which has a little prior information.

Comparison (RSW)

- Reference Speaker Weighting (RSW)
  - The new speaker's model is a linear combination of the reference models.

    $$m = w(1)m(1) + \cdots + w(R)m(R)$$
  - The rest part is same as eigenvoice method.
  - Good with medium or large-vocabulary systems.
    - For class p and speaker r

      $$\mu_m^{(s)}(p, r) = c(p, r) + \nu_m^{(s)}$$
    - $m(r) = [c(1, r), c(2, r), \cdots, c(P, r)]$ speaker dependent
    - *v* is speaker independent
    - For new speaker model, from *m*, the vector *m(S)* is obtained by means of the ML. (equivalent to finding the weights)
    - As the number R of reference speakers grows, it becomes more expensive in terms of memory and computation.

Comparison (Vowel Classification)

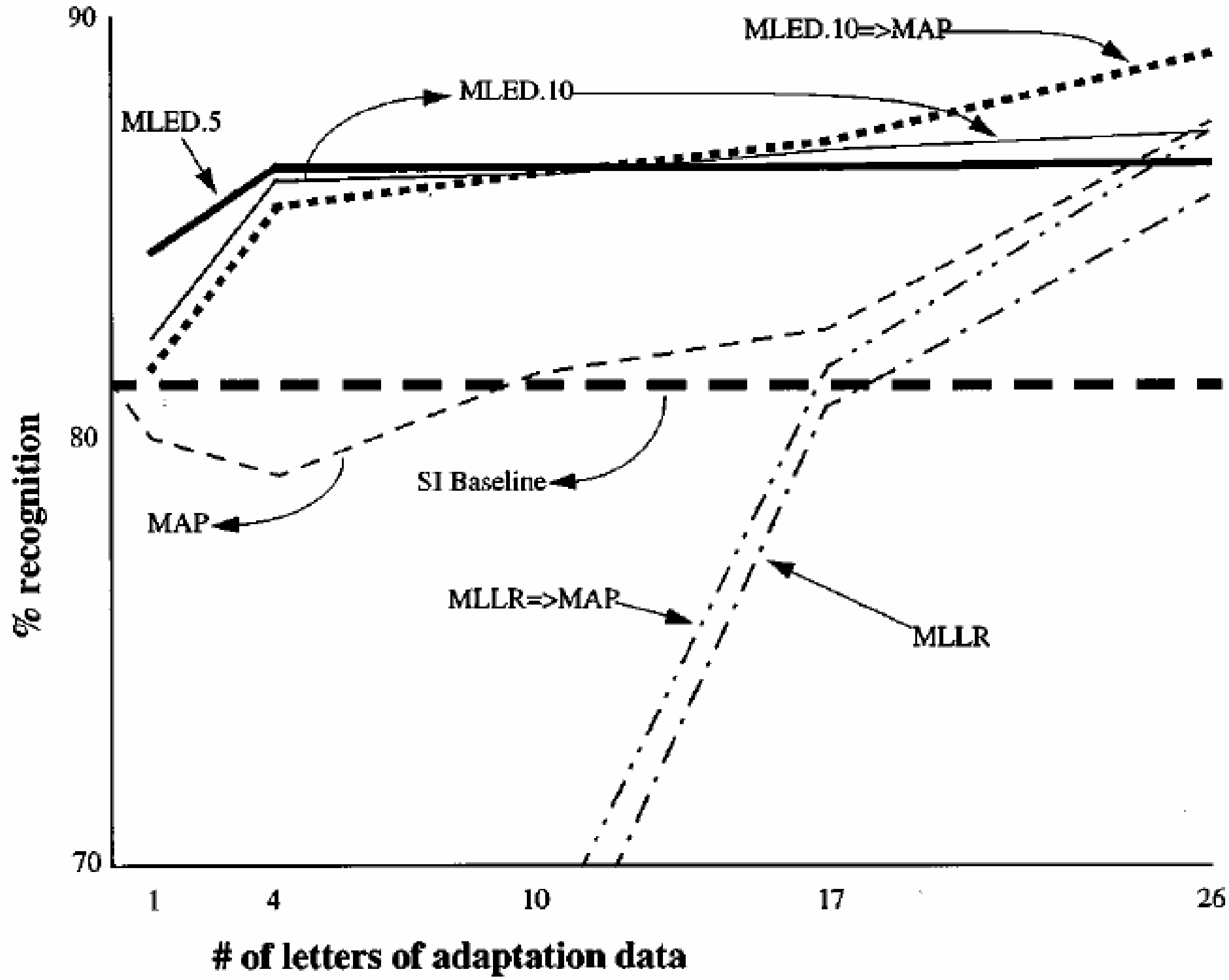
- Vowel Classification
  - PCA to the parameters of a vowel classifier (MoG)

Experiments

- Database
  - R = 120 (reference speakers) and 30 test speakers
  - D = 2808 (parameter dimension)
    - = 26 characters x 6 states (single Gaussian) x 18 features
  - PCA (0.,,, K eigenvoices) and MLED (2 iterations)

- In figure
  - MLED.5 : MLED with K=5
  - MLED.10=>MAP : MAP with MLED.10 model as prior
  - Otherwise, SI is prior.

Experiments - Accuracy



Experiments – Accuracy

MLED.5 Recognition Rate (1 Letter of Adaptation Data)

RANK	LETTER	% CORRECT	RANK	LETTER	% CORRECT
1	'V'	85.7	10	'L'	84.0
2	'D'	85.6	10	'X'	84.0
2	'T'	85.6	10	'Y'	84.0
3	'G'	85.5	11	'F'	83.9
3	'J'	85.5	11	'I'	83.9
4	'C'	85.3	11	'S'	83.9
4	'E'	85.3	12	'N'	83.8
5	'A'	85.2	13	'U'	83.7
5	'B'	85.2	14	'Q'	83.5
6	'P'	85.0	15	'R'	83.2
7	'H'	84.7	16	'M'	82.8
8	'K'	84.6	16	'O'	82.8
9	'Z'	84.4	17	'W'	82.2

Robustness

- Sensitiveness to changes in the reference speaker models.

MLED—CHANGING TRAINING DATA QUANTITY

Type	2 prod, 120 spkrs	2 prod, 60 spkrs	2 prod, 30 spkrs	1 prod, 120 spkrs
MLED.1	85.0	82.0	81.5	84.7
MLED.5	87.1	86.1	85.4	86.2
MLED.10	88.1	86.3	85.6	87.5

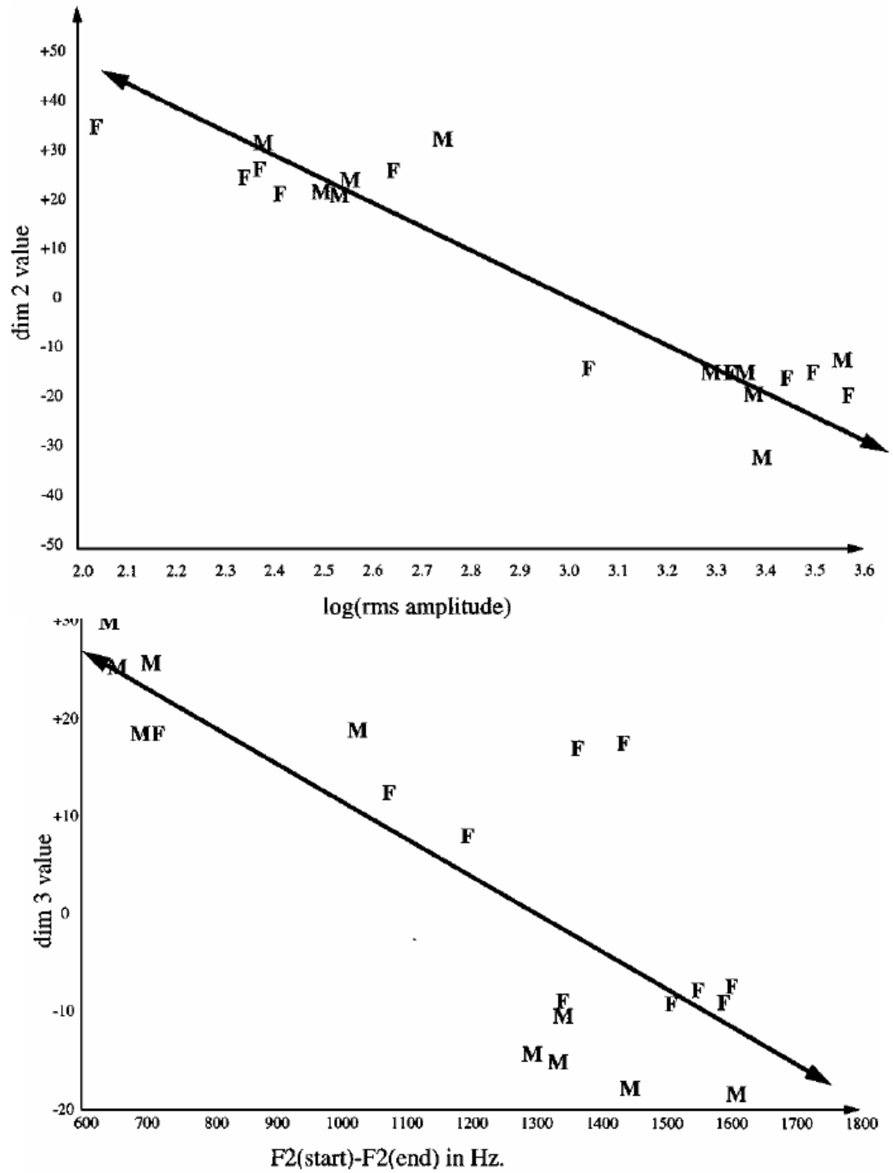
ADAPTIVE TRAINING EXPERIMENTS

Type	Training	Full	balanced 17 letters	random (17-let. average)
MLED.1	ML	85.0	81.8	84.3
MLED.1	adaptive	84.9	84.1	84.2
MLED.5	ML	87.1	81.0	85.6
MLED.5	adaptive	87.4	86.1	85.9
MLED.10	ML	88.1	81.0	85.9
MLED.10	adaptive	88.0	86.1	86.6

Interpretation

- "it is by no means universally true." – I. T. Jolliffe

- 1st eigenvector :
  - sex (strong)
- 2nd eigenvector :
  - amplitude (strong)
- 3rd eigenvector :
  - second-formant (maybe)
- 4th eigenvector :
  - changes in pitch (maybe)



Future work

- Extensions of the Eigenvoice Approach
  - Hybrid
    - such as MLED+MAP in Fig. 2. (Done)
    - How about allowing K to rise as the data increases?
    - MLED + MLLR
  - Discriminative training of the reference SD models;
  - Environment adaptation
  - LDA rather than PCA
  - Learning basis vectors by ML rather than PCA
  - Eigenvoice adaptation of state transition probabilities and Gaussian standard deviations
    - I don't think so. There are definitely some correlations between states and Gaussian, though.

Future work

- Training models for Large-Vocabulary Systems
  - There will be insufficient data per reference speaker.
  - The computational and storage requirements of this naïve extention of small-vocabulary methodology would be onerous.

  - Principals
    - Inter-speaker variability (in K-space)
    - Intra-speaker variability (by the Gaussians)
    - Efficient pooling of training data from different speakers
      - Remove speaker-dependent characteristics from training data at the beginning of training, rather than based on SI model.

Summary

- Fast way to speaker adaptation based on Eigenvoices.
- Maximum Likelihood Eigen-Decomposition
  - Better than MLLR and MAP.
- For large vocabulary data, MLED+MAP is best.