Aalto University
School of Science
and Technology

Eigenvoices

A review of eigenvoice adaptation techniques for ASR

Peter Smit

peter.smit@tkk.fi

Speech Group
Adaptive Informatics Research Centre

November 29, 2010

Outline

Eigenvoices Introduction

Reduction Techniques

- PCA
- Maximum Likelihood Eigenspace

Decomposition Techniques

- Projection
- Maximum Likelihood Eigen Decomposition

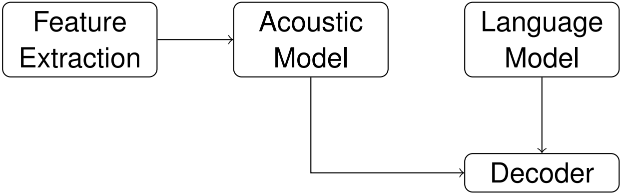
Combination of Eigenvoices and other Adaptation techniques

- MLLR in space estimation
- Use Eigenvoices to restrict MLLR parameter freedom
- Maximum a Posteriori Linear Regression

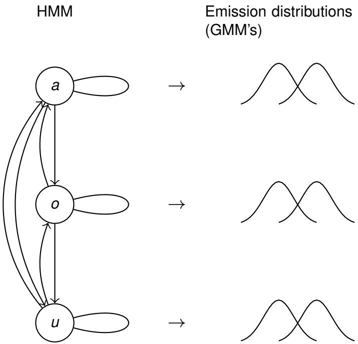
Other applications of eigenvoices

Discussion and Conclusion

ASR system


Feature Extraction
Acoustic Model
Language Model
Decoder

Acoustic Model



Acoustic Model adaptation

3 classes of adaptation techniques for Acoustic Models[1]

- Maximum Likelihood Linear Regression (MLLR / CMLLR)
- Maximum a posteri parameter estimation (MAP)
- Clustering / Eigenvoices

[1] P.C. Woodland. “Speaker adaptation for continuous density HMMs: A review”. In: *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*. 2001. URL: http://www.isca-speech.org/archive/adaptation/adap_011.html.

Eigenvoices techniques

Eigenvoices were first proposed by Kuhn, 1998[2] and were inspired by the eigenfaces technique.

- Find a low-dimensional space describing speaker variability
- Find a mapping between every point in this space and a Speaker Dependent model
- Find a method of mapping new speech to a point in this space

[2] R. Kuhn et al. "Eigenvoices for speaker adaptation". In: *Fifth International Conference on Spoken Language Processing*. 1998. URL: `http://www.isca-speech.org/archive/icslp_1998/i98_0303.html`.

Creating a low-dimensional space of speakers

- Train multiple Speaker Dependent models
- Apply a dimensionality reduction technique (DRT) on the parameter vectors of these models
- DRT should provide a mapping back to the SD-model space

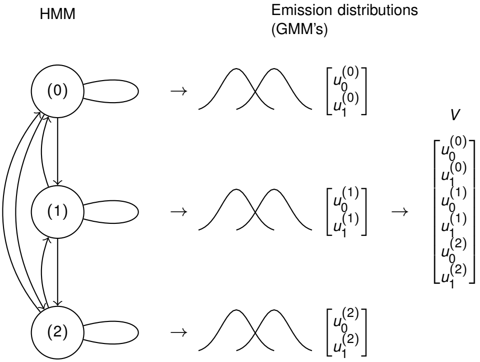
Eigenvoice Decomposition

- Find the coefficients that work best with the adaptation data
- Construct a new SD-model from the eigenvoice coefficients

Why so vague? Where are the formulas?

- A lot of different options!
- Reduction: (kernel-)PCA (LSES), MLES, LDA, ICA, ...
- Decompostion: Projection, MLED, ...

Create single vector from SD-model



PCA

- Apply PCA on supervectors made from SD-model
- First *n* eigenvectors are the bases of new space
- The coefficients **w** can be used to construct a new supervector: $V' = \mathbf{w}^T E$
- Covariances of new SD models are equal to SI model covariances

- Costly / impossible for big LVCSR models

Maximum Likelihood Eigenspace

- Integrated estimation of coefficients and eigenvoices as hidden data in Baum-Welch training procedure[3]

$$\hat{M} = \arg\max_{M} \sum_{q=1}^{T} \int \log L(O, w|M) P_0(w, q) dw$$

- $P_0(w, q)$ can be used as superviced prior (for example dialect or sex)

[3] Patrick Nguyen, Christian Wellekens, and Jean-Claude Junqua. "Maximum likelihood eigenspace and MLLR for s recognition in noisy environments". In: *Sixth European Conference on Speech Communication and Technology*. Cites 1999, pp. 2519–2522.

Maximum Likelihood Eigenspace cont.

- The seed eigenvoices have to be found with PCA or LDA
- Possible for big models, main cost is memory usage in training
- Worse performance than PCA technique

Projection

- Train a SD-model from the adaptation data and construct V
- Find coefficients P in eigenspace: $P = E \times E^T \times V$
- Construct new SD-model from P: $V' = P^T \times E$

Problem:

Projection

- Train a SD-model from the adaptation data and construct V
- Find coefficients P in eigenspace: $P = E \times E^T \times V$
- Construct new SD-model from P: $V' = P^T \times E$

Problem:

- Every HMM-state must be present in adaptation data to be able to have an estimate for all parameters

Maximum Likelihood Eigen Decomposition

- Find the coefficients that maximize the likelihood of the adaptation data
- $\arg\max_{\mathbf{w}} P(O|\lambda, \mathbf{w})$
- Very low computational cost, only requirement is to solve k equations with k variables where k is the dimension of the coefficients

This method is the “standard” decomposition technique

MLED - The Math

The auxiliary function $Q(\lambda, \hat{\lambda})$ is defined by

$$Q(\lambda, \hat{\lambda}) = -\frac{1}{2} P(O \mid \lambda) \sum_s \sum_t \gamma_s(t) \mathbf{f}(\mathbf{o}_t, s)$$

with

$$\mathbf{f}(\mathbf{o}_t, s) = \left(n \log(2\pi) + \log |C_s| + (\mathbf{o}_t - \hat{\mu}_s)^T C_s^{-1} (\mathbf{o}_t - \hat{\mu}_s) \right)$$

Analytical solution:

$$\sum_t \sum_s \gamma_s(t) \mathbf{e}_s^T C_s^{-1} \mathbf{o}_t = \mathbf{w}^T \sum_t \sum_s \gamma_s(t) \mathbf{e}_s^T C_s^{-1} \mathbf{e}_s$$

Using MLLR in space estimation

The normal method needs to train a lot of separate Speaker Dependent models to estimate space. MLLR comes to the rescue...

- Use MLLR transformed models as “SD-models”
- Saves the effort of training separate SD-models –> computationally cheaper
- Regression Classes can be used

Creating a space based on MLLR transformations

- Instead of building the space on the model parameters, build it on (C)MLLR transformation parameters
- More efficient training of space (lower number of parameters)
- Makes the coefficients transferable between models

Use Eigenvoices to restrict MLLR parameter freedom

Chen et al. 2000[4] proposes a method to combine eigenvoices and MLLR

- Use eigenvoices as a 'prior' to MLLR
- Eigenvoice adaptation restricts the parameters of the MLLR matrix, therefore giving a better MLLR transform for little adaptation data. The restricting is done by 'smoothing' the parameters

[4] K. Chen et al. "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression". In: *Sixth International Conference on Spoken Language Processing*. Citeseer. 2000.

Maximum a Posteriori Linear Regression

MLED+MAP was already proposed in Kuhn, 1998[1] and Maximum a Posteriori Linear Regression was proposed in Chen, 2001[5]

- Both: Use eigenvoices as a prior for MAP adaptation
- Chen: Use Probabilistic PCA as reduction technique and incorporate the estimation of weights and MAP estimation
- Very usable for bigger amounts of adaptation data, but because of the good prior also performing well for small amounts of data

[5] K.T. Chen and H.M. Wang. "Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptat In: *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*. Vol. 1. Citesee 2001.

Speech Separation

In Weiss and Ellis, 2010[6] eigenvoices are used in a speech separation algorithm.

- Low number of parameters made easier separation possible
- The eigenvoice parameters were easily adjustable for gain variability.

[6] R.J. Weiss and D.P.W. Ellis. "Speech separation using speaker-adapted eigenvoice speech models". In: *Computer Speech & Language* 24.1 (2010), pp. 16–29.

Speech Synthesis

Toda, 2006[7] used eigenvoices in Speech Synthesis

- Use eigenvoice parameters for voice conversion

[7] T. Toda, Y. Ohtani, and K. Shikano. “Eigenvoice conversion based on Gaussian mixture model”. In: *Ninth Internat* *Conference on Spoken Language Processing*. Citeseer. 2006.

Discussion

Advantages of Eigenvoices

- Works well with amounts of adaptation data
- Online calculations are computationally very cheap

Disadvantages of Eigenvoices

- Offline calculations can be very expensive for big models
- Limited use when there is a lot of adaptation data
- No real standardization in techniques

Conclusion

Eigenvoice techniques are useful when

- There is only very little adaptation data
- In combination with other adaptation techniques

Questions?

- Questions?