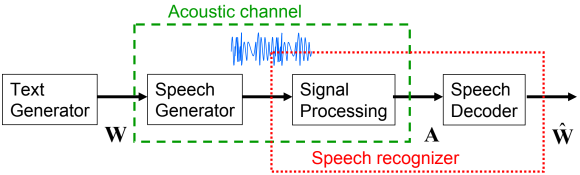
Linear Transforms in Automatic Speech Recognition: Estimation Procedures & Integration of Diverse Acoustic Data

Stavros Tsakalidis
Advisor: Prof. William Byrne
Thesis Committee: Prof. William Byrne, Prof. Sanjeev Khudanpur, Prof. Paul Sotiriadis, Prof. Trac Tran

Center for Language and Speech Processing and
Department of Electrical and Computer Engineering
The Johns Hopkins University

September 23, 2005

Introduction



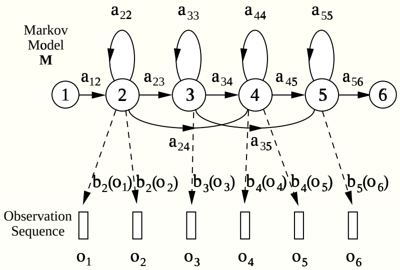
Source-channel model of speech recognition

- Purpose of speech recognition is to find an estimate $\hat{W}$ of W
- *Maximum a posteriori* (MAP) decoder:

$$\hat{W} = \operatorname*{argmax}_{W'} P(W'|A) = \operatorname*{argmax}_{W'} P(W')P(A|W')$$

  - To compute $P(W)$ we use a statistical *language model* $\tilde{P}(W)$
  - To compute $P(A|W)$ we use a statistical *acoustic model* $P_\theta(A|W)$
- We will investigate the use of linear transforms in acoustic models to
  - improve acoustic modeling
  - work with diverse acoustic data

Acoustic Model



- The Hidden Markov Model (HMM) is the standard acoustic model
- Output emission density is modeled as a Gaussian : $b_s(o_\tau;\theta) = \mathcal{N}(o_\tau;\mu_s,\Sigma_s)$
- Applying affine transform $T = [b\ A]$ to the observation vector $\zeta = [1\ o']'$ $(Ao + b = T\zeta)$ the Gaussian density is reparameterized as:

$$b_s(\zeta;\theta) = |A| \cdot \mathcal{N}(T\zeta;\mu_s,\Sigma_s) = \frac{|A|}{\sqrt{(2\pi)^m|\Sigma_s|}} e^{-\frac{1}{2}(T\zeta-\mu_s)'\Sigma_s^{-1}(T\zeta-\mu_s)}$$

Motivation

- Most ASR systems
  - HMM-based acoustic models
  - Gaussian state-dependent emission densities
    - Diagonal covariance assumption
  - Trained from homogeneous data collections similar to the recognition task
  - Parameters estimated via Maximum Likelihood or discriminative criteria
- Linear transforms in ASR
  - Feature-space :
    - Maximum Likelihood Linear Transform (MLLT)
    - Heteroscedastic Linear Discriminant Analysis (HLDA)
  - Model-space :
    - Maximum Likelihood Linear Regression (MLLR)

Goals

- Discriminative estimation of feature-based transforms
  - Estimate transforms under discriminative criteria rather than ML
  - Incorporate discriminative transforms in discriminative HMM training
- Maximum-A-Posteriori (MAP) estimation of feature-based transforms
  - Estimate transforms under MAP criterion
  - Useful for dealing with problems posed by sparse training data
- Develop acoustic normalization procedures for the integration of diverse acoustic data
  - Enlarge acoustic training set with speech data that would otherwise lead to performance degradation

Outline

Outline

Discriminative Estimation of Linear Transforms

- Focus on MLLT technique:
  - Transforms acoustic data to ease diagonal covariance assumption
  - Estimation of transforms via ML criterion
- Discriminative training of HMMs useful in LVCSR tasks
- Success triggered an interest in the use of discriminative linear transforms
  - *Maximum Mutual Information Linear Regression*
  - *Conditional Maximum Likelihood Linear Regression*
  - Both techniques use model-space linear transforms
- Prior Work: MLLT was developed as ML technique, but has also been used with MMI
  - The AT&T LVCSR-2001 system used:
    - feature-based transforms obtained by ML estimation techniques, and
    - were then fixed throughout the subsequent iterations of MMI model estimation
- Goal: Estimation of feature-based transforms under the Conditional Maximum Likelihood (CML) criterion

Conditional Maximum Likelihood Auxiliary Function

- CML criterion uses a general auxiliary function similar to EM

$$\bar{\theta} : \sum_{s} \sum_{\tau} \left(\gamma_{s,\tau}(\theta) - \gamma^{g}_{s,\tau}(\theta)\right) \cdot \nabla_{\theta} \log q(o_\tau | s; \bar{\theta}) + \sum_{s} D_s \int q(o|s;\theta) \nabla_{\theta} \log q(o|s;\bar{\theta}) do = 0$$

- State dependent distributions are reparametrized to incorporate linear transforms
- CML version of MLLT is readily obtained
- Goal is to maximize $P(W|A;\theta)$ by alternatively updating the transforms and HMM parameters
- As a result both transforms and HMM parameters are estimated discriminatively

Discriminative Likelihood Linear Transforms

- Apply affine transform to the observation vector : $T\zeta = Ao + b$
- Under the preceding model, emission density of state s becomes

$$q(\zeta|s;\theta) = \frac{|A_{\mathcal{R}(s)}|}{\sqrt{(2\pi)^m|\Sigma_s|}} e^{-\frac{1}{2}(T_{\mathcal{R}(s)}\zeta-\mu_s)^T\Sigma_s^{-1}(T_{\mathcal{R}(s)}\zeta-\mu_s)}$$

Objective: We estimate transforms & HMM parameters under the CML criterion

- The transforms obtained under this criterion are termed

Discriminative Likelihood Linear Transforms (DLLT)

- This estimation is performed as a two-stage iterative procedure:
  a) Maximize CML criterion with respect to the transforms while keeping the Gaussian parameters fixed
  b) Compute the Gaussian parameters using the updated values of the transforms

DLLT Estimation

As in MLLT (Gales '97), the i^{th} row of the transformation matrix is found by

$$[\bar{T}_r]_i = (\alpha p_i + k_{r,i})\, G_{r,i}^{-1}$$

where

$$G_{r,i} = \sum_{s:\mathcal{R}(s)=r} \frac{1}{\sigma_{s,i}^2} \left(\sum_{\tau} \left(\gamma_{s,\tau}(\theta) - \gamma_{s,\tau}^{g}(\theta) \right) \hat{\zeta}_\tau \hat{\zeta}_\tau^T + D_s J_s \right)$$

$$k_{r,i} = \sum_{s:\mathcal{R}(s)=r} \frac{\mu_{s,i}}{\sigma_{s,i}^2} \left(\sum_{\tau} \left(\gamma_{s,\tau}(\theta) - \gamma_{s,\tau}^{g}(\theta) \right) \hat{\zeta}_\tau^T + D_s [J_s]_1 \right)$$

$$J_s = \begin{bmatrix} 1 & (A_r^{-1}(\mu_s - b_r))^T \\ A_r^{-1}(\mu_s - b_r) & A_r^{-1}(\Sigma_s + (\mu_s - b_r)(\mu_s - b_r)^T) A_r^{-1\,T} \end{bmatrix}.$$

DLLT: System Description

Task: English conversational telephone speech transcription

- Acoustic Training/Test Set
  - Training: 16.4 Hrs Switchboard-1 & 0.5 Hr CallHome
  - Test: 866 utts 2000 Switchboard-1 evaluation (SWBD1) & 913 utts 1998 Switchboard-2 evaluation (SWBD2)
  - 2001 JHU LVCSR minitrain/minitest system
- Acoustic Model
  - Standard HTK flat-start training procedure
  - Tied state, cross-word, context-dependent triphones
  - 4000 unique triphone states
  - 6 mixtures per speech state
- Language Model
  - 33k-word trigram language model
- Fixed number (450) of linear transforms

DLLT Results: ML vs. CML transforms

Goal: Compare ML trained transforms to CML trained transforms

- Gaussian parameters fixed throughout transform updates
- Test whether CML transforms improve over ML transforms
- Validate CML as a modeling procedure

Transform reestimation only		
	Word Error Rate (%)	
	SWBD1	SWBD2
ML	41.1	51.1
ML+MLLT	39.1	50.3
ML+DLLT	38.5	49.7

- CML transforms outperform ML transforms

DLLT Results: Fully discriminative training

Goal: Investigate fully discriminative training compared to ML training

A

DLLT	Word Error Rate (%)	
	SWBD1	SWBD2
ML	41.1	51.1
DLLT-1it	38.2	49.2
DLLT-2it	37.3	48.9

B

MLLT and DLLT	Word Error Rate (%)	
	SWBD1	SWBD2
ML	41.1	51.1
MLLT-1it	38.4	49.6
MLLT-2it	38.2	49.5
MLLT-6it	37.8	49.0
MLLT+DLLT-1it	37.1	48.4

- DLLT outperforms MLLT
- DLLT is additive with MLLT
- DLLT works best when initialized by MLLT

DLLT Summary

- Developed discriminative estimation procedures for feature-based transforms
- Integrated discriminative linear transforms into CML estimation for LVCSR
- Discriminative versions outperform ML training
- Validated CML as a modeling procedure
- DLLT used in the JHU LVCSR-2002 Evaluation system with similar success
- Discriminative linear transforms became standard in state-of-the-art ASR systems

Outline

Maximum-A-Posteriori Estimation of Linear Transforms

- Provides a way of incorporating prior information about the transformation parameters in the training process
- MAP estimate of transform T:

$$\bar{T} = \arg\max_{T} p(\hat{o}_1^{\hat{l}} | M_{\hat{w}_1^{\hat{n}}}, T; \theta)\ p(T)$$

- MAP estimate can be formulated in an EM algorithm

$$R(\bar{T}, T) = Q(\bar{T}, T) + \log p(\bar{T})$$

where $Q(\bar{T}, T)$ is the conventional ML auxiliary function

Selection of the prior density

- MAP solution is strongly related to
  a) the choice of the prior density and
  b) the specification of the parameters for the prior densities
- Use the matrix variate normal distribution:

$$p(T;\eta) \propto |\Phi|^{-(m+1)/2}|\Psi|^{-m/2}\exp\{-\frac{1}{2}\operatorname{tr}\left(\Phi^{-1}(T-\Lambda)\Psi^{-1}(T-\Lambda)^T\right)\}$$

where $T,\Lambda \in \mathbb{R}^{m\times(m+1)}$, $\Phi \in \mathbb{R}^{m\times m}, \Phi > 0$, $\Psi \in \mathbb{R}^{(m+1)\times(m+1)}, \Psi > 0$.

1. Closed-form solution for the transformation parameters
2. Related to the location-scale family of distributions
   - Λ is the location parameter and Ψ, Φ are the scale parameters

MAP Feature-Space Transforms

As in MLLT and DLLT, the i^{th} row of the transformation matrix is found by

$$[\bar{T}]_i = \left(\tilde{\alpha} p_i + \tilde{k}_i\right) \tilde{G}_i^{-1}$$

where

$$\tilde{G}_i = G_i + \frac{1}{2}(\phi^{-1})_{i,i}\left(\Psi^{-1} + \Psi^{-T}\right)$$

and

$$\tilde{k}_i = k_i + \frac{1}{2}[\Phi^{-1}\Lambda\Psi^{-1}]_i + \frac{1}{2}[\Phi^{-T}\Lambda\Psi^{-T}]_i - \frac{1}{2}\sum_{j\neq i}[T]_j\left((\phi^{-1})_{i,j}\Psi^{-1} + (\phi^{-1})_{j,i}\Psi^{-T}\right)$$

- The MAP and ML update equations are similar, except for the additional terms in the definition of $\tilde{G}_i$, $\tilde{k}_i$ and $\tilde{\alpha}$.

Outline

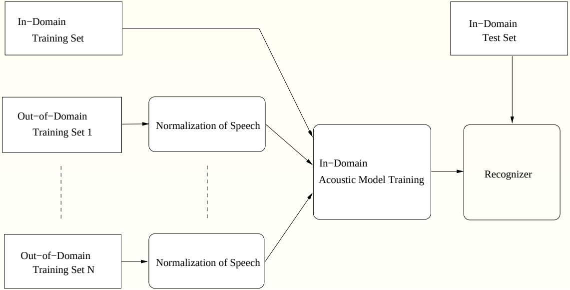
Motivation

- State-of-the-art ASR systems require the collection of huge amounts of transcribed speech for the task which the ASR system will be applied to
- Collection and transcription of speech is expensive and time-consuming
- Available speech collections come from a variety of sources
- Great challenge: Integration of heterogeneous data sources in acoustic training

Integration of Diverse Acoustic Data

- Speech collections differ due to a wide range of factors:
  - Acoustic channel, sampling rate
  - Speaking style, speaker age and education
  - Domain or topic, dialect, language
- Incorporating diverse acoustic data may lead to performance degradation
- Objective: Develop acoustic normalization techniques that:
  - Augment acoustic training set with out-of-domain data

Acoustic Normalization Overview



- Map the out-of-domain feature space onto the in-domain data
- Employ linear transforms
- Obtain a larger in-domain training set by transforming the out-of-domain data
- Performance measured by improvements on the in-domain test set

Cross-Corpus Acoustic Normalization

- Collection of C training sets, indexed by c
  - $c = t$ denotes the in-domain data set
- Assume in-domain training data sufficient representative of in-domain test data
- Transforms estimated over the out-of-domain training sets
- Emission density of state s is reparameterized as

$$q(\zeta|s, c; \theta) = \frac{|A^{(c)}|}{\sqrt{(2\pi)^m|\Sigma_s|}} e^{-\frac{1}{2}(T^{(c)}\zeta - \mu_s)' \Sigma_s^{-1} (T^{(c)}\zeta - \mu_s)}$$

- Distribution depends on c; the training set to which is applied
- Assume in-domain data does not need to be normalized at the corpus level :

$$A^{(t)} = \mathbf{I} \text{ and } b^{(t)} = \mathbf{0}$$

Cross-Corpus Parameter Estimation with Normalization

- Incorporate source identity into the observed random process

$$(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{l}}, \hat{c}_1^{\hat{l}})$$

- Goal: Estimate transforms and HMM parameters via ML

$$\theta^* = \operatorname*{argmax}_{\theta} p(\hat{o}_1^{\hat{l}} | M_{\hat{w}_1^{\hat{n}}}, \hat{c}_1^{\hat{l}}; \theta)$$

- This estimation is performed as a two-stage iterative procedure:
  - Maximize ML criterion with respect to the transforms while keeping the Gaussian parameters fixed
  - Compute the Gaussian parameters using the updated values of the transforms

Acoustic Corpora Description

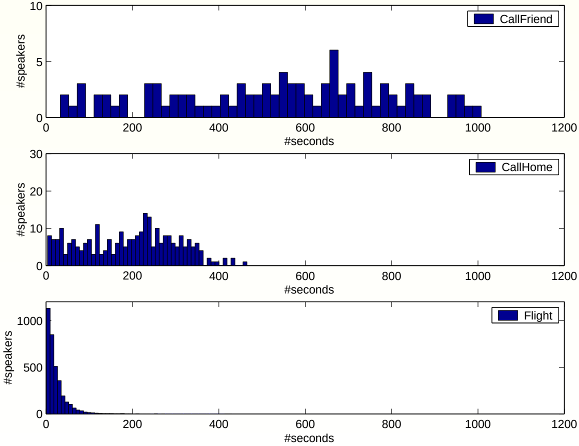
Task: Mandarin conversational telephone speech transcription for the CallFriend domain

Available acoustic corpora: Three conversational Mandarin databases collected over the telephone

	Data Sources in Training	#Hours	#Conversations
In-domain	CallFriend (CF)	14	42
Out-of-domain	CallHome (CH)	14	100
	Flight (FL)	22	1800

- CallFriend : both parties located in USA and Canada
- CallHome : calls originated in N. America but placed overseas
- Flight : based entirely in China
- Speakers in CallFriend & CallHome took advantage of a free phone call
- Flight : conversations between travel agents and customers about flight enquiries and reservations

Acoustic Corpora Analysis



- Majority of the speakers in the Flight corpus have limited training data!

System Description

- Baseline Acoustic Models
  - Standard HTK flat-start training procedure
  - Number of unique states depends on the amount of training data
- Language Model (LM)
  - Three bigram LMs trained from each set of transcriptions
  - Linearly interpolated with weight chosen to minimize perplexity on held-out CallFriend transcriptions

Corpus	CF	CH	FL
Weight	0.77	0.21	0.02

  - This bigram used for *all* decoding experiments

Unormalized Out-of-Domain Acoustic Data

Data Sources in Training				Character Error Rate (%)	
CF	CH	FL	Total Hours	SI	SI+MLLR
$\surd$			14	60.8	58.7
	$\surd$		14	62.2	59.8
		$\surd$	22	69.2	65.8
$\surd$	$\surd$		28	57.9	55.9
$\surd$		$\surd$	36	60.8	58.7
$\surd$	$\surd$	$\surd$	50	59.3	56.6

- CallHome comparable to CallFriend
- Flight significantly worse than CallFriend and CallHome
- Adding Flight to CallFriend & CallHome degrades performance !

Normalized Out-of-Domain Acoustic Data

Data Sources & Normalization				Character Error Rate (%)	
CF	CH	FL	#transforms	SI	SI+MLLR
I	I			57.9	55.9
I	T		1 per corpus	57.6	55.8
I	I	I		59.3	56.6
I	I	T	1 per corpus	58.1	55.7
I	T	T	1 per corpus	57.8	55.5

- 'T' / 'I' indicates that a source was included in training with / without normalization, resp.
- 1 (global) transform per corpus
- Cross-corpus normalization prevents degradation in performance

Speaker-to-Corpus Normalized Out-of-Domain Acoustic Data

Data Sources & Normalization				Estimation Criterion	CER	
CF	CH	FL	#transforms		SI	SI+MLLR
I	T	T	1 per corpus	ML	57.8	55.5
I	T	T	1 per speaker	ML	57.7	55.2
I	T	T	1 per speaker	MAP	57.6	54.9

- A significant amount of speakers in the out-of-domain corpora have limited amounts of training data.
- MAP estimation procedure reduced the overfitting of the speaker-level transforms.

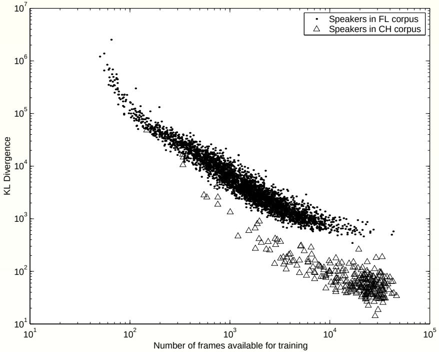
Distance Measures Between Model Sets

- Objective: Investigate how 'close' are the MAP derived transforms $T_{(MAP)}$ to the ML derived transforms $T_{(ML)}$
- Two sets of parameters: $\eta_{(MAP)} = (\theta, T_{(MAP)})$ and $\eta_{(ML)} = (\theta, T_{(ML)})$
- Use an average KL divergence between pairs of Gaussians on a per mixture component level, defined as

$$D(T_{(MAP)}, T_{(ML)}) = \frac{1}{\sum_{s=1}^{S} M(s)} \sum_{s=1}^{S} \sum_{m=1}^{M(s)} D_{KL}\left(q(\zeta|s, m; T_{(MAP)}), q(\zeta|s, m; T_{(ML)})\right)$$

- $M(s)$ is the number of mixture components in state s

Average Kullback-Leibler divergence $D(T^k_{(MAP)}, T^k_{(ML)})$



- KL divergence $D(T^k_{(MAP)}, T^k_{(ML)})$ is 'on-the-average' inversely proportional to the amount of the training data

Speaker Adaptive Training on Normalized Out-of-Domain Acoustic Data

Data Sources & Normalization				Character Error Rate (%) on CF	
CF	CH	FL	#transforms	SI+MLLR	SAT+MLLR
I	I	I		56.6	55.6
I	T	T	1 per corpus	55.5	54.6

- We compare the performance of SAT acoustic models trained over:
  - Unnormalized acoustic data
  - An in-domain training set created by transforming the out-of-domain corpora prior to SAT.
- Gains from SAT and cross-corpus normalization are almost exactly additive, and are thus capturing complementary influences
- Results show that SAT can be further improved by 1.0% if we first compensate for the cross-corpus differences across the training sets.

Cross-Corpus Normalization Summary

- Simple attempts to add out-of-domain data degrade performance
- Developed acoustic normalization procedure for enlarging acoustic training set with out-of-domain data, using techniques developed for speaker normalization in training
- Experimental results show that cross-corpus acoustic normalization makes it possible to combine severely mismatched corpora to increase the amount of available training data
- Cross-domain normalization can also improve Speaker Adaptive Training.
  - Experimental results show that performing SAT over cross-corpus normalized data effectively doubles the gains obtained from SAT alone on this corpus.
  - Cross-corpus normalization and SAT procedures yield additive improvement.

Outline

Highlights of Thesis Contributions

- Developed a novel estimation procedure that finds Discriminative Linear Transforms (DLLT) jointly with MMI for feature normalization
  - Result: DLLT provides a discriminative estimation framework that outperforms its Maximum Likelihood counterpart (MLLT)
- Developed a structural maximum a posteriori estimation framework for feature-space transforms.
  - Result: MAP estimation of linear transforms provides improved performance relative to ML estimation procedures when data sparseness was encountered.
- Described a cross-corpus acoustic normalization procedure for enlarging an ASR acoustic training set with out-of-domain acoustic data.
  - Result: Improve performance by adding a severely mismatched corpus that otherwise leads to performance degradation.
  - Result: Cross-corpus normalization can also improve Speaker Adaptive Training (SAT).
    - Performing SAT over cross-corpus normalized data yields additive improvement.

Outline

Future Work

- **Combination of discriminative training within a Bayesian formalism**
  - To improve MMI estimates, based on insufficient data, incorporate prior information about the transformation parameters into the discriminative training framework
- **Controlling the contribution of out-of-domain data**
  - Linear transforms attempt to capture non-linguistic acoustic variations between heterogeneous collections.
  - Mismatch between acoustic sources may be nonlinear
  - As an alternative, a weighted version of the classical ML estimation framework can be used
  - We expect the primary challenge to be
    1. the identification of the out-of-domain data that needs to be downweighted
    2. the estimation of the corresponding weights
  - One possibility is to use a distance that measures the deviation of the normalized observation from the current estimated model