MPhil in Machine Learning and Machine Intelligence

Module MLMI2: Speech Recognition

L13: Discriminative Sequence Training

Phil Woodland
pcw@eng.cam.ac.uk

Michaelmas 2021

Cambridge University Engineering Department

Introduction

These lectures concentrate on the use of alternative **discriminative sequence training** objective functions for both GMM-HMM systems and DNN-HMM systems.

- basic theory and implementation is developed for GMM-HMMs
- then applied to DNN-HMMs also

We will include for GMM-HMMs

- Motivation for discriminative training
- Maximum Mutual Information (MMI) Estimation & Issues for LVCSR
- Lattice based approach
- Minimum Phone Error (MPE)
- Prior information: I-smoothing and discriminative MAP

Then will discuss discriminative sequence training for DNN-HMMs

- Gradient computations
- Optimisation approaches for DNN-HMMs
- Performance
- Lattice Free MMI

Discriminative Training for GMM-HMMs

- Standard GMM-HMM training uses **maximum likelihood** estimation (MLE)
- MLE optimisation criteria is

$$\mathcal{F}_{\text{MLE}}(\lambda) = \sum_{r=1}^{R} \log p_\lambda\left(\boldsymbol{O}_r|\mathcal{M}_{\mathcal{H}_r}\right)$$

  $\mathcal{H}_r$ is the (correct) transcription for utterance r and $\mathcal{M}_{\mathcal{H}_r}$ the corresponding model.
- MLE be **optimal** if several unrealistic assumptions met, including
  - Infinite training set size
  - Model correctness
- Neither condition met for speech recognition!
- hence interesting to investigate alternatives, especially **discriminative** schemes. In particular we will look at:
  - Maximum Mutual Information Estimation (MMIE)
  - Minimum Phone Error (MPE)

MMIE Basics

MMIE maximises the sentence level posterior (Bahl et al., 1986) : in log form

$$\mathcal{F}_{\text{MMIE}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda\left(\boldsymbol{O}_r|\mathcal{M}_{\mathcal{H}_r}\right) P\left(\mathcal{H}_r\right)}{\sum_{\mathcal{H}} p_\lambda\left(\boldsymbol{O}_r|\mathcal{M}_{\mathcal{H}}\right) P\left(\mathcal{H}\right)}$$

- **Numerator** is likelihood of data given correct transcription (as for MLE)
- **Denominator** expands total likelihood in terms of **all** word sequences
- Can compute denominator by finding likelihood through composite HMM with all recognition constraints (recognition model)
- Maximise **ratio** of numerator (MLE term) to denominator
- More closely related to **word error rate** than MLE

Strictly Conditional Maximum Likelihood Estimator at utterance level

- but here MMI since LM fixed

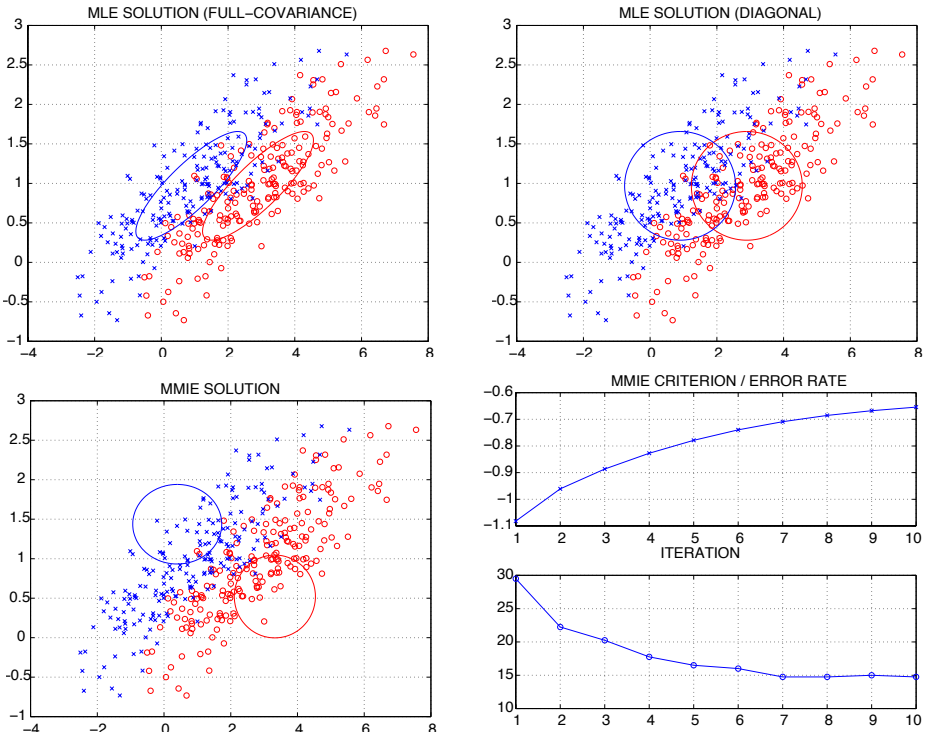
MMIE weights training data **unequally** (well classified small weight)

- MLE gives all training samples equal weight
- Sensitive to **outliers**
- Use of an error measure instead of MMIE would reduce sensitivity

MMIE was used in the early 1990's for small vocabulary work, but at the time it was thought not applicable for large vocabularies.

Simple MMIE Example: incorrect model assumptions

- 2-d data from full covariance Gaussian: Modelled with diagonal covariance Gaussian



MMIE Issues for LVCSR

Need to have effective **optimisation technique** that scales well to large systems.

- Extended Baum-Welch for GMM-HMMs (Gopalakrishnan et al., Normandin, 1991)

$$\hat{\mu}_{jm} = \frac{\left\{\theta_{jm}^{\text{num}}(\mathcal{O}) - \theta_{jm}^{\text{den}}(\mathcal{O})\right\} + D\mu_{jm}}{\left\{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}}\right\} + D}$$

$$\hat{\sigma}_{jm}^2 = \frac{\left\{\theta_{jm}^{\text{num}}(\mathcal{O}^2) - \theta_{jm}^{\text{den}}(\mathcal{O}^2)\right\} + D(\sigma_{jm}^2 + \mu_{jm}^2)}{\left\{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}}\right\} + D} - \hat{\mu}_{jm}^2$$

In the above:

- Gaussian occupancies (summed over time) are γ_{jm}.
- $\theta_{jm}(\mathcal{O})$ and $\theta_{jm}(\mathcal{O}^2)$ are sums of data and squared data respectively, weighted by occupancy.
- num and den denote correct word sequence, & recognition model respectively.

Denominator requires computation of all sentence likelihoods: approximate with **lattices** (Valtchev et al, 1996)

Require good **generalisation**

- Can reduce training set error rate: need to reduce test-set errors!
- Need to keep gains with large numbers of parameters

Improve generalisation increasing **"confusable"** data for training

- Use **acoustic scaling** to broaden posterior distribution across denominator (Woodland & Povey, 2002).
- Typically scale by the inverse of the language model scale factor i.e. HMM likelihood and LM probability as:

$$\frac{1}{s}\log p(\mathcal{O}|\mathcal{H}) + \log P(\mathcal{H})$$

  where s is the normal language model scale factor.
- This scaling has no effect if finding the **1-best** but greatly modifies the estimated **posterior** probability of confusable words. [Note same scaling used in e.g. confusion network construction].
- **Weakened language model** to increase focus on acoustics (Schluter et al., 1999). Typically use a unigram or a very small bigram model.

Lattice Based MMIE

Typical Lattice-Based approach is as follows:

- Use a **word-lattice** to represent **numerator** & **denominator** terms
- Recognise every training sentence with MLE models (denominator)
- Generate phone-marked lattices with **model boundary** times (for efficiency & scaling)
- Accumulate statistics for EBW via forward-backward pass on lattice
- "Exact match" lattice search
  - Only run forward-backward between boundaries: more efficient
  - Use **acoustic scaling of complete segments** (by inverse of LM scale factor)
- Run EBW algorithm for several iterations
- F-B passes uses unigram (or v. small bigram) language model probabilities
- Parameter updates
  - Standard updating formulae for means/variances
  - Gaussian specific D constant with flooring
- Can optionally re-generate lattices part-way through training
  - Normally works well with just the MLE generated models

NAB/WSJ MMIE Results: GMM-HMMs

- Standard HTK large vocab GMM-HMM LVCSR system (no adaptation, no HLDA)
- 66 hour training set

#Mix Comp	H1 dev		H1 eval	
	MLE	MMIE	MLE	MMIE
1	13.64	11.36	15.64	13.16
2	11.84	10.12	13.19	11.31
4	10.67	9.42	11.25	10.59
12	9.30	8.80	9.96	9.40

- Bigger reductions in WER for **simpler systems**
  - in general larger gains as amount of data per parameter increases
- All model complexities **reduce WERs** with MMIE training (unlike initial attempts at large vocabulary discriminative training ...)

If train on WSJ-type data and test on broadcast news data

- MMIE always better than MLE even with severe mismatch!
- Advantage larger for larger mismatch

Incorporating Prior Information in Discriminative Training

Use of discriminative criteria can easily cause over-training

- Can modify the so-called **"weak sense auxiliary function"** associated with Extended Baum-Welch algorithm with additional smoothing terms

Want to make normal discriminative training **more robust** when there is insufficient data

- ML estimate requires less data to accurately estimate than MMIE

The ML estimate of the parameter values can be used as the centre of an appropriately defined prior distribution for each Gaussian

- This yields **I-Smoothing**

$$\mu_j = \frac{\{\theta_j^{\text{num}}(\mathcal{O}) - \theta_j^{\text{den}}(\mathcal{O})\} + D_j\hat{\mu}_j + \tau^I\mu_j^{\text{ml}}}{\{\gamma_j^{\text{num}} - \gamma_j^{\text{den}}\} + D_j + \tau^I}$$

- τ^I determines influence of "prior" (ML estimate) on the final MMIE estimate.

MPE Objective Function

- **Minimum Phone Error** (Povey & Woodland, 2002) minimises the following **minimum Bayes' Risk** formulation:

$$\mathcal{F}_{\text{MPE}}(\lambda) = \sum_{r}^{R} \sum_{\mathcal{H}} P(\mathcal{H}|\mathcal{O}_r; \mathcal{M}) \mathcal{L}(\mathcal{H}, \hat{\mathcal{H}}_r)$$

- $\mathcal{L}(\mathcal{H}, \hat{\mathcal{H}})$ is loss function between hypothesis $\mathcal{H}$ and reference $\hat{\mathcal{H}}$
- For MPE, loss function if is number of phone errors in hypothesis $\mathcal{H}$
- $\mathcal{F}_{\text{MPE}}(\lambda)$ is weighted average of the loss function over all $\mathcal{H}$
  - MPE is smoothed approx to **phone error in a word recognition context**
- Minimum Word Error (MWE) just counts errors differently (word errors in $\mathcal{H}$)
- Error measure reduces sensitivity to outliers
- Can use lattice-based implementation (requires time-based alignments for errors) and new statistics computation to still use EBW update formulae
- MPE and MWE train to minimise the **Bayes' Risk** with particular loss functions
- I-smoothing essential for MPE (& helps a little for MMI)

MMIE/MPE on Conversational Telephone Speech: GMM-HMMs

	% WER Train	% WER eval98	% WER abs redn (test)
MLE baseline	47.2	45.6	–
MMIE (τ=0)	37.7	41.8	3.8%
MMIE (τ=200)	35.8	41.2	4.2%
MPE (τ=100)	34.4	40.8	4.8%

GMM-HMMs trained on 265hr training set. Train is lattice unigram

- MPE/I-smoothing gives around 1% abs lower WER than MMIE results
- For large data-sets, further small improvements of typically 0.4% by using a dynamic MMI "prior" (rather than a dynamic ML prior) for MPE training
- Similar improvements from MPE have been found for many tasks/languages.
- Gains from discriminative training **increase** for
  - **Simpler** models: very useful for small footprint systems
  - **Larger** training sets give more improvements over MLE (results up to 2,000 hours of training data in Evermann et al, 2005)
- Discriminative training is more **robust** to training/test mismatch

Discriminative MAP Adaptation

MAP is a useful scheme when a significant amount of adaptation data available:

- increasing adaptation data tends to Maximum Likelihood estimation: called **ML-MAP**

For **discriminative MAP** schemes (Povey et al, 2003):

- increasing adaptation data tends to discriminative estimation;
- maximum mutual information (**MMI-MAP**) and minimum phone error (**MPE-MAP**) forms

For adaptation/task porting (small amount of task-specific data), ML estimate not robust

- use a ML-MAP estimate as the prior
- Use count-smoothing ML-MAP with prior parameters ($\tilde{\mu}_j$)

$$\mu_j = \frac{\{\theta_j^{\text{num}}(\mathcal{O}) - \theta_j^{\text{den}}(\mathcal{O})\} + D_j\hat{\mu}_j + \tau^{\prime}\mu_j^{\text{map}}}{\{\gamma_j^{\text{num}} - \gamma_j^{\text{den}}\} + D_j + \tau^{\prime}}$$

where $\mu_j^{\text{map}} = \frac{\theta_j^{\text{num}}(\mathcal{O}) + \tau\tilde{\mu}_j}{\gamma_j^{\text{num}} + \tau}$

Two smoothing variables for MMI-MAP

- τ determines how "close" the prior is to the ML estimate
- $\tau^{\prime}$ determines how much the prior influences the final estimate.
- Similar form may be used for MPE-MAP.

Discriminative Linear-Transform Based Adaptation also possible (supervised/unsupervised).

Large Margin Training & BMMI

- Also interest in using **large margin** techniques to ensure good **generalisation**
- Basic idea is to ensure that for each training example

$$\log p_\lambda(\mathcal{O}_r, \mathcal{H}_r) - \log p_\lambda(\mathcal{O}_r, \mathcal{H}) \geq \rho, \qquad \forall \mathcal{H} \neq \mathcal{H}_r$$

- Can make margin, ρ, dependent on the difference between the reference and closest competitor (**dynamic margin**)
  - Measured using phone error, word error, frame-level etc.
- Want to integrate into lattice-based discriminative training approach
- Boosted MMI (Povey et al, 2008) is one form

$$\mathcal{F}_{\text{BOOSTMMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\mathcal{O}_r|\mathcal{M}_{\mathcal{H}_r})P(\mathcal{H}_r)}{\sum_{\mathcal{H}} p_\lambda(\mathcal{O}_r|\mathcal{M}_{\mathcal{H}})P(\mathcal{H})e^{\beta D(\mathcal{H},\mathcal{H}_r)}}$$

- β is a boosting parameter; $D(\mathcal{H}, \mathcal{H}_r)$ measures the difference between the correct and each of the denominator lattices
  - $D(\mathcal{H}, \mathcal{H}_r)$ could be the **phone error rate** or some other error difference
- Focus training effort on poorly recognised, low likelihood alternatives by making $\beta > 0$
- If use phone error, can sometimes get small reductions in WER over standard MPE

Discriminative Sequence Training for DNN-HMMs

Previously seen that DNN-HMMs can be trained effectively using the cross-entropy criteria.

- **frame-based** discriminative classification (conditional maximum likelihood).
- requires frame-based labels

In speech recognition, we are actually interested in the performance of systems for **sequence-level discrimination**, such as WER

- Individual word labels have equal weight: not individual frames

To improve sequence-level discrimination, can use the same type of objective functions as discriminative training of GMM-HMMs!

- MMI: utterance level conditional maximum likelihood
- MPE (i.e. phone level minimum Bayes' risk)
- state-level MBR: referred to as sMBR
  (computed at the context-dependent clustered state level).

May wish to combine sequence and frame-based objective functions.

- Could be done by interpolating e.g. the MPE criterion with the CE model (**F-smoothing**).

Computational Approach

In the same way as for GMM-HMMs, need to compute information:

- From training utterances using statistics derived from a full recognition network
- Approximate recognition network the full recognition using lattices
- Compute lattices on the training set using previously trained set of CE DNNs

Using these lattices, and starting from the CE-trained model, go through the training set and

- Run lattice-forward backward for each utterance and calculate relevant gradients for error back propagation for DNN parameters
- Accumulate statistics for each utterance/set of utterances
- Update DNN-parameters

Note that we might perform multiple epochs of e.g. MPE training with the same lattices or alternatively re-generate the lattices with models.

The above was the standard approach for DNN-HMMs, but more recently there has been interest in **Lattice-Free MMI** (Povey et al, 2016)

- Doesn't require explicit lattice generation
- To make this efficient enough on current GPUs, model structure needs to be significantly altered and simplified

Objective Function Gradients

Recall that for the CE objective function

$$\mathcal{F}_{\text{CE}} = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_k^{(n)} \log y_k^{(n)}$$

where the $t_k^{(n)}$ values are targets for pattern n and $y_k^{(n)}$ is the value of the output node k.

For CE with softmax output layer found the partial differential of the objective function wrt the input to output node k, a_k:

$$\frac{\partial \mathcal{F}_{\text{CE}}}{\partial a_k} = \sum_{n=1}^{N} y_k^{(n)} - t_k^{(n)}$$

For discriminative criteria we again need to find the differentials of the objective function.

For MMI for a particular frame n of an utterance:

$$\frac{\partial \mathcal{F}_{\text{MMI}}}{\partial a_k} = \kappa(\gamma_k^{\text{num}}(n) - \gamma_k^{\text{den}}(n))$$

where κ is the acoustic scaling factor; $\gamma_k^{\text{num}}(n)$ is posterior prob of state k in numerator lattice and $\gamma_k^{\text{den}}(n)$ is posterior prob of state k in denominator lattice.

For MBR objective functions can derive a suitable gradient based on the **MBR "posterior"**.

Optimisation Methods

The discriminative objective functions require a forward-backward pass through the lattice for each utterance in order to find the posterior values.

Can be updated using SGD as for CE training, but only **randomised at the utterance level**

- This means that a small learning rate is needed
- Update after each utterance processed
- Often DNN forward-backward computation on GPU and lattice forward-backward on CPU

Alternative optimisation strategies are possible. In particular can:

- Use **larger batches** of utterances (could compute statistics in parallel)
- Perform far **fewer updates** than normal SGD: more work per update, slower (but better?) convergence.

E.g. 2nd order method called **Hessian Free training** (Kingsbury, 2009; Kingsbury et al., 2012)

- uses a Gauss-Newton approximation of the Hessian
- uses linear conjugate gradient method to avoid explicit matrix inverse

Alternative is **Natural Gradient**. Can be used with large batches (Haider & Woodland, 2017)

- modifies standard gradient with inverse of empirical Fisher information (FI) matrix
- FI can be approximated as the outer product of the Jacobian of the MMI criterion
- can be more effective than HF and SGD
- can also be combined with HF (Haider et al, 2021).

Performance

The WER reduction (WERR) from discriminative sequence training over CE can be from almost zero to about 15% WERR.

Actual improvement depends on

- task
- model type (more advanced models may be more difficult)
- activation function (ReLU style activations harder to get improvements)
- optimisation method

Effect of DNN sequence training on MGB data (Woodland et al, 2015):

- Hybrid DNNs use 5 layers of 1000 sigmoid nodes and 6000 output targets.
- 40 FBANK inputs context window 9 frames.
- For tandem, bottleneck DNNs of 39 dim. inserted into same config. (but with ReLU)
- Tandem GMMs model 78 features (incl. PLP+HLDA).
- 700h training set from distributed data, manual segmentation, 64k vocab, 4-gram LM.

AM	%WER
GMM-HMM ML HLDA	42.7
GMM-HMM MPE	40.7
DNN-HMM Hybrid CE	28.4
Tandem SI MPE	27.0
DNN-HMM Hybrid MPE	25.9

In the above example, using SGD, discriminative sequence training (MPE) for hybrid DNN-HMMs gives a further **9% WERR over CE**.

Lattice Free MMI

Lattice-free MMI (Povey et al, 2016) allows training with the MMI sequence-level discriminative objective function **without first generating utterance-specific denominator word lattices**.

- Aim to be trained without an initial CE model from a random initialisation (although some versions use F-smoothing)
- To be efficient on GPUs, simplify model structure

Simplifications:

- Use context dependent HMMs (reduced minimum dwell time) with std decision trees
- Three-fold reduced frame rate for modelling (but use all 10 ms frames in TDNN structure)
- Fixed phone-level language model for denominator (e.g. phone trigram with no back-off)
- Max utterance length (segment utterances)
- Other timing constraints from numerator alignment
- fully GPU-based computation

This has been extended to an **end-to-end** version (Hadian et al, 2018) with some modifications

- monophone, full-biphone or character based
- no decision trees
- no alignment of numerator required
- completely trained from flat-start

LF-MMI approaches now widely used and **often gives leading hybrid ASR WERs**.

Summary

Sequence discriminative training is **effective** for both GMM-HMM and DNN-HMM systems.

Important to address basic issues:

- lattice-based statistics calculation
- generalisation (acoustic scaling, weakened LMs)

Interesting properties for GMM-HMMs

- WER difference to ML is bigger with more data (more effective with fewer parameters)
- Improvements under within-task and cross-task conditions

Also can be applied to adaptation

- Discriminative MAP adaptation schemes (better task porting)
- Discriminative linear transforms (supervised and unsupervised adaptation)

Minimum Phone Error training more effective than MMIE

- Example of Minimum Bayes' Risk training

State-of-the-art DNN-HMMs also use discriminative sequence training

- Same lattice based approach (but note recent trend for lattice-free MMI)
- Start with CE trained model
- same MBR-based objective functions (MPE, sMBR) and MMI
- use SGD or other optimisation approaches (HF and NG)
- trend towards using lattice-free MMI to train from scratch (& possibly free of decision trees and dictionary)

HTK 3.5 supports lattice-based MPE and MMI for both GMM-HMMs and DNN-HMMs (SGD).

Discriminative Sequence Training References

- L. Bahl, P. Brown, P.V de Souza, P.V. & R. Mercer. "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition." *Proc. ICASSP*, 1986.
- G. Evermann, H.Y. Chan, M.J.F. Gales, B. Jia, D. Mrva, P.C. Woodland & K. Yu. "Training LVCSR Systems on Thousands of Hours of Data". *Proc. ICASSP*, 2005.
- P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo, & A. Nadas. "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems." *IEEE Trans. Information Theory*, vol. 37, 1991.
- H. Hadian, H. Sameti, D. Povey & S. Khudanpur, "End-to-end speech recognition using lattice-free MMI". *Proc. Interspeech*, 2018.
- M.A. Haider & P.C. Woodland. "Sequence Training of DNN Acoustic Models With Natural Gradient". *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017.
- M.A. Haider, C. Zhang, F.L. Kreyssig, P.C. Woodland. "A Distributed Optimisation Framework Combining Natural Gradient with Hessian-Free for Discriminative Sequence Training". *Neural Networks*, 2021.
- B. Kingsbury. "Lattice-based Optimization of Sequence Classification Criteria for Neural-Network Acoustic Modeling." *Proc. ICASSP*, 2009.
- B. Kingsbury, T.N. Sainath & H. Soltau, "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization," *Proc. Interspeech*, 2012.
- Y. Normandin, D. Morgera. "An Improved MMIE Training Algorithm for Speaker-Independent Small Vocabulary Continuous Speech Recognition." *Proc. ICASSP* 1991.
- D. Povey & P.C. Woodland. "Minimum Phone Error and I-Smoothing for Improved Discriminative Training." *Proc. ICASSP*, 2002.
- D. Povey, P.C. Woodland & M.J.F. Gales. "Discriminative MAP for Acoustic Model Adaptation." *Proc. ICASSP*, 2003.

Discriminative Sequence Training References (ctd.)

- D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, & K. Visweswariah. "Boosted MMI for model and feature-space discriminative training." *Proc. ICASSP*, 2008.
- D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, & S. Khudanpur. "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI." *Proc. Interspeech*, 2016.
- T.N. Sainath, B. Kingsbury & H. Soltau, "Optimization Techniques to Improve Training Speed of Deep Neural Networks for Large Speech Tasks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2267–2276, 2013.
- R. Schlüter, B. Müller, F. Wessel, & H. Ney. "Interdependence of Language Models and Discriminative Training." *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 1999
- V. Valtchev, J.J. Odell, P.C. Woodland & S.J. Young. "MMIE Training of Large Vocabulary Recognition Systems." *Speech Communication*, Vol. 22, pp. 303–314, 1997.
- K. Veselý, A. Ghoshal, L. Burget, D. Povey. "Sequence-Discriminative Training of Deep Neural Networks." *Proc. Interspeech*, 2013.
- L. Wang & P.C. Woodland. "MPE-based Discriminative Linear Transforms for Speaker Adaptation." *Computer Speech and Language*, vol. 22. pp. 256–272, 2008.
- P.C Woodland & D. Povey. Large scale discriminative training of hidden Markov models for speech recognition *Computer Speech & Language*, Vol. 16, pp. 25–47, 2002.
- P.C. Woodland, X. Lui, Y. Qian, C. Zhang, M.J.F. Gales, P. Karanasou, P. Lanchantin & L. Wang. "Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* 2015.