MMI-MAP and MPE-MAP for Acoustic Model Adaptation

D. Povey, M.J.F. Gales, D.Y. Kim, & P.C. Woodland

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
{dp10006,mjfg,dyk21,pcw}@eng.cam.ac.uk

Abstract

This paper investigates the use of discriminative schemes based on the maximum mutual information (MMI) and minimum phone error (MPE) objective functions for both task and gender adaptation. A method for incorporating prior information into the discriminative training framework is described. If an appropriate form of prior distribution is used, then this may be implemented by simply altering the values of the counts used for parameter estimation. The prior distribution can be based around maximum likelihood parameter estimates, giving a technique known as I-smoothing, or for adaptation it can be based around a MAP estimate of the ML parameters, leading to MMI-MAP, or MPE-MAP. MMI-MAP is shown to be effective for task adaptation, where data from one task (Voicemail) is used to adapt a HMM set trained on another task (Switchboard). MPE-MAP is shown to be effective for generating gender-dependent models for Broadcast News transcription.

1. Introduction

In recent years the use of discriminative training techniques such as Maximum Mutual Information Estimation (MMIE) have been shown to outperform conventional Maximum Likelihood Estimation (MLE) for large vocabulary HMM-based speech recognition [8]. However adaptation techniques for these models such are still generally based on MLE: for instance, Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori (MAP) adaptation. While it has been shown that MLLR can be effective for speaker adaptation of MMI-trained models [8], and that conventional MAP can be effective for task adaptation of MMI-trained models [1], it is interesting to investigate if there are additional benefits from the use of discriminative objective functions in adaptation. Previous work in discriminative adaptation includes a MAP-type scheme described in [4] and discriminative transform estimation [7].

This paper describes a framework, originally discussed in [6], for incorporating prior information into the estimation of model parameters via the use of *weak-sense* auxiliary functions. Using the appropriate prior distribution, the MAP adaptation for standard MLE (ML-MAP) may be viewed as simple count smoothing in contrast to the standard MAP scheme described in [2]. Furthermore, using weak-sense auxiliary functions it is simple to extend the MAP scheme to incorporate discriminative training criteria. This again results in smoothing the usual discriminative update counts with the prior counts.

The paper is arranged as follows. In Section 2 the concept of weak-sense auxiliary functions are described. Section 3 describes how prior information can be incorporated into the parameter estimation and describes specific discriminative MAP schemes. Section 4 presents the experimental results.

This work was funded by the European Commission under the Language project Le-5 Coretex. Extensive use was made of equipment donated by IBM under an SUR award.

2. Weak-Sense Auxiliary Functions

The discriminative MAP procedures used in this paper are derived using weak-sense auxiliary function [6]. The theory behind the use of these functions is described in the next section. It is then shown how it may be applied to MMI training.

2.1. Strong- and Weak-Sense Auxiliary Functions

In [6] strong-sense and weak-sense auxiliary functions were described. The attributes of these functions are briefly summarised below. In this paper $\hat{\lambda}$ is used to represent the current model parameters and λ the parameters to be estimated.

- **Strong-sense** auxiliary function: a function $\mathcal{G}(\lambda,\hat{\lambda})$ is a strong-sense auxiliary function for a function $\mathcal{F}(\lambda)$ *around* $\hat{\lambda}$, if

$$\mathcal{G}(\lambda,\hat{\lambda}) - \mathcal{G}(\hat{\lambda},\hat{\lambda}) \leq \mathcal{F}(\lambda) - \mathcal{F}(\hat{\lambda}), \qquad (1)$$

  where $\mathcal{G}(\lambda,\hat{\lambda})$ is a smooth function of λ. This is the standard form of auxiliary function used in expectation maximisation. Maximisation of the auxiliary is guaranteed to not decrease the value of $\mathcal{F}(\lambda)$, and hence iterative use of auxiliary functions around each new parameter estimate will find a local maximum of the function.

- **Weak-sense** auxiliary function: a function $\mathcal{G}(\lambda,\lambda')$ is a weak-sense auxiliary function for a function $\mathcal{F}(\lambda)$ *around* $\hat{\lambda}$, if

$$\left.\frac{\partial}{\partial\lambda}\mathcal{G}(\lambda,\hat{\lambda})\right|_{\lambda=\hat{\lambda}} = \left.\frac{\partial}{\partial\lambda}\mathcal{F}(\lambda)\right|_{\lambda=\hat{\lambda}}. \qquad (2)$$

  The condition of being a weak-sense auxiliary function can be considered a minimum condition for an auxiliary function to be useful for optimisation. If the objective function has a maximum at $\hat{\lambda}$, the weak-sense auxiliary function is also bound to have its maximum at $\hat{\lambda}$. However, in contrast to the strong-sense auxiliary function increasing the value of the weak-sense auxiliary does not necessarily increase the value of the original.

Despite the limitations of weak-sense auxiliary functions compared to strong-sense functions,there are advantages to their use. The primary advantage is that a weak-sense function may be specified for many situations where strong-sense functions cannot be used. As weak-sense auxiliary functions do not guarantee an increase in the original function, they are comparable to standard gradient descent techniques. However, the advantage of using a weak-sense auxiliary function is that there is no need to determine the appropriate learning rate, or use second-order statistics. The weak-sense auxiliary function may be selected so that it has a simple closed-form for the parameter estimation. Normally these will need to be smoothed in some form to try to ensure that the value of the original function increases.

There are thus two functional forms to select when using weak-sense auxiliary functions. First the auxiliary function of the function to be optimised is required. Second an appropriate form of smoothing function is required; it must be some function with its maximum at $\hat{\lambda}$.

2.2. Weak-sense auxiliary functions for MMIE

This section describes how a weak-sense auxiliary function may be used to optimise the MMI criterion for training HMMs and how, given the appropriate smoothing function, it yields the standard extended Baum-Welch (EBW) update rules. Considering only a single training utterance, $\mathcal{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_T\}$ and using a fixed language model[1], the MMI criterion may be expressed as

$$\mathcal{F}(\lambda) = \log p(\mathcal{O}|\mathcal{M}^{\text{num}}) - \log p(\mathcal{O}|\mathcal{M}^{\text{den}}) \tag{3}$$

where $\mathcal{M}^{\text{num}}$ and $\mathcal{M}^{\text{den}}$ are HMMs corresponding to the correct transcription (numerator term) and all possible transcriptions (denominator term) respectively. It is not possible to define a strong-sense auxiliary function for this expression, since the second term is negative. Therefore the inequality of equation (1) will no longer hold. However, it is possible to linearly combine individual weak-sense auxiliary functions to form an overall weak-sense auxiliary function, even when there is negation.

As a strong-sense auxiliary function is by definition also a weak-sense auxiliary function, it is natural to use the standard strong-sense auxiliary function associated with ML estimation as an appropriate form for the weak-sense auxiliary function. Thus a possible weak-sense auxiliary function for the numerator term (considering a single Gaussian per state with a single dimension) is

$$\begin{aligned}\mathcal{G}^{\text{num}}(\lambda, \hat{\lambda}) &= \sum_{t=1}^{T}\sum_{j=1}^{J} \gamma_j^{\text{num}}(t) \log\left(p_\lambda(o_t|s_j)\right) \\ &= \sum_{j=1}^{J} \mathcal{Q}(\gamma_j^{\text{num}}, \theta_j^{\text{num}}(\mathcal{O}), \theta_j^{\text{num}}(\mathcal{O}^2), \lambda_j) \quad (4)\end{aligned}$$

where $\lambda_j = \{\mu_j, \sigma_j^2\}$,

$$\begin{aligned}&\mathcal{Q}(\gamma_j, \theta_j(\mathcal{O}), \theta_j(\mathcal{O}^2), \lambda_j) = \\ &\quad -\frac{1}{2}\left(\gamma_j \log(2\pi\sigma^2) + \frac{\theta_j(\mathcal{O}^2) - 2\theta_j(\mathcal{O})\mu_j + \gamma_j\mu_j^2}{\sigma_j^2}\right) \quad (5)\end{aligned}$$

s_j indicates state j of the system, $\gamma_j(t)$ is the posterior probability of being in state s_j at time t given $\hat{\lambda}$, and the sufficient statistics to evaluate the function for the numerator are given by $\theta_j^{\text{num}}(\mathcal{O}) = \sum_{t=1}^{T} \gamma_j^{\text{num}}(t) o_t$, $\theta_j^{\text{num}}(\mathcal{O}^2) = \sum_{t=1}^{T} \gamma_j^{\text{num}}(t) o_t^2$ and $\gamma_j^{\text{num}} = \sum_{t=1}^{T} \gamma_j(t)$ the occupancy of the state. Similarly the auxiliary function for the denominator term alone can be defined. These two may then be combined to yield a candidate weak-sense auxiliary function for the MMI criterion.

[1] This is sometimes known as conditional maximum likelihood training.

As previously mentioned, in order to improve stability of the training process, a smoothing function, $\mathcal{G}^{\text{sm}}(\lambda, \hat{\lambda})$, can be added. This may be any function with a zero differential w.r.t. λ around the current estimate $\lambda = \hat{\lambda}$. As such combining this with any weak-sense auxiliary will still be a valid weak-sense auxiliary function. Hence, for MMIE the complete weak sense auxiliary function will have the form

$$\mathcal{G}^{\text{mmi}}(\lambda, \hat{\lambda}) = \mathcal{G}^{\text{num}}(\lambda, \hat{\lambda}) - \mathcal{G}^{\text{den}}(\lambda, \hat{\lambda}) + \mathcal{G}^{\text{sm}}(\lambda, \hat{\lambda}). \tag{6}$$

One possible form for $\mathcal{G}^{\text{sm}}(\lambda, \hat{\lambda})$ is to use D_j "effective" observations which yield the current state parameters, $\hat{\lambda}$, as the ML estimate, thus automatically satisfying the requirements for the smoothing function. This may be written in the same form as equation (4)

$$\mathcal{G}^{\text{sm}}(\lambda, \hat{\lambda}) = \sum_{j=1}^{J} \mathcal{Q}(D_j, D_j\hat{\mu}_j, D_j(\hat{\mu}_j^2 + \hat{\sigma}_j^2), \lambda_j), \tag{7}$$

where D_j are positive smoothing constants for each state j. The above analysis can be simply extended for multiple Gaussian components per state.

Optimising the weak-sense auxiliary function simply requires combining the sufficient statistics for each of the individual auxiliary functions. The global maximum of $\mathcal{G}^{\text{mmi}}(\lambda, \hat{\lambda})$ for the mean and variance of component m of state j are given by

$$\mu_{jm} = \frac{\left\{\theta_{jm}^{\text{num}}(\mathcal{O}) - \theta_{jm}^{\text{den}}(\mathcal{O})\right\} + D_{jm}\hat{\mu}_{jm}}{\left\{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}}\right\} + D_{jm}} \tag{8}$$

$$\sigma_{jm}^2 = \frac{\left\{\theta_{jm}^{\text{num}}(\mathcal{O}^2) - \theta_{jm}^{\text{den}}(\mathcal{O}^2)\right\} + D_{jm}(\hat{\sigma}_{jm}^2 + \hat{\mu}_{jm}^2)}{\left\{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}}\right\} + D_{jm}} - \mu_{jm}^2 \tag{9}$$

where D_{jm} is set on a per-Gaussian level as described in [8] and determines the convergence-rate and stability of the update rule. These are the standard update rules obtained from the extended Baum-Welch (EBW) algorithm [3], though derived using weak-sense auxiliary functions. Similarly, update equations may also be derived for the component priors and transition probabilities.

3. Incorporating Prior Information

In this section the incorporation of a prior into the weak-sense auxiliary function framework is discussed. The derivation of I-smoothing and discriminative MAP based on MMI (MMI-MAP) and MPE (MPE-MAP) is described.

By definition, any function is both a weak and strong-sense auxiliary function of itself around any point. Thus it is possible to add any form of log prior distribution over the model parameters to a weak-sense auxiliary function and still have a weak-sense auxiliary function for a MAP version of the original function. Adding a log-prior to the MMI criterion yields

$$\mathcal{F}(\lambda) = \log p(\mathcal{O}|\mathcal{M}^{\text{num}}) - \log p(\mathcal{O}|\mathcal{M}^{\text{den}}) + \log p(\lambda) \tag{10}$$

The extra term can be directly added to the associated weak-sense auxiliary function leading to

$$\mathcal{G}(\lambda, \hat{\lambda}) = \mathcal{G}^{\text{mmi}}(\lambda, \hat{\lambda}) + \log p(\lambda). \tag{11}$$

The exact form of the log-prior distribution affects the nature of the MAP update. One of the major issues, and choices, in MAP estimation is how to obtain this prior distribution.

3.1. I-smoothing

I-smoothing for discriminative training [5] may be regarded as the use of a prior over the parameters of each Gaussian, with the prior being based on the ML statistics. The log prior likelihood is defined as

$$\log p(\lambda_{jm}) = \mathcal{Q}\left(\tau^I, \tau^I \frac{\theta^{\rm num}_{jm}(\mathcal{O})}{\gamma^{\rm num}_{jm}}, \tau^I \frac{\theta^{\rm num}_{jm}(\mathcal{O}^2)}{\gamma^{\rm num}_{jm}}, \lambda_{jm}\right) \quad (12)$$

This log-prior is the log-likelihood of τ^I points of data with mean and variance equal to the numerator (correct model) mean and variance. The MMIE update formula for the mean is then

$$\mu_{jm} = \frac{\{\theta^{\rm num}_{jm}(\mathcal{O}) - \theta^{\rm den}_{jm}(\mathcal{O})\} + D_{jm}\hat{\mu}_{jm} + \tau^I \mu^{\rm ml}_{jm}}{\{\gamma^{\rm num}_{jm} - \gamma^{\rm den}_{jm}\} + D_{jm} + \tau^I} \quad (13)$$

where $\mu^{\rm ml}_{jm} = \frac{\theta^{\rm num}_{jm}(\mathcal{O})}{\gamma^{\rm num}_{jm}}$.

I-smoothing can also be directly implemented by altering the numerator statistics [6]. A similar form of prior with MPE training yields I-smoothing for MPE.

3.2. MMI-MAP

In the context of adapting a HMM set, the use of ML statistics accumulated from the adaptation data as the center of the prior may not be robust since there may not be enough data to estimate the ML Gaussian parameters. In this case it is preferable to estimate the center of the prior in a fashion similar to standard ML-MAP. The technique denoted MMI-MAP is the use of ML-MAP estimates of the Gaussian parameters to estimate the centre of a prior used to smooth the MMI-trained parameters. MMI-MAP has two distinct levels of operation.

In the first level of MAP the unadapted mean and variance $\tilde{\mu}_{jm}$ and $\tilde{\sigma}_{jm}$ are used as the prior, and the numerator (ML) statistics as the adaptation data. The parameters are effectively estimated by using count smoothing, related to the weak-sense auxiliary functions described here, rather than the ML-MAP described in [2]. The expressions for the ML-MAP mean and variance are:

$$\mu^{\rm map}_{jm} = \frac{\theta^{\rm num}_{jm}(\mathcal{O}) + \tau\tilde{\mu}_{jm}}{\gamma^{\rm num}_{jm} + \tau} \quad (14)$$

$$\sigma^{\rm map\,2}_{jm} = \frac{\theta^{\rm num}_{jm}(\mathcal{O}^2) + \tau(\tilde{\mu}^2_{jm} + \tilde{\sigma}^2_{jm})}{\gamma^{\rm num}_{jm} + \tau} - \mu^{\rm map\,2}_{jm}. \quad (15)$$

The ML-MAP parameters are then used to generate the prior for the second level of MMI-MAP. The count weighting for this prior is set using an additional variable τ^I. The estimate of the MMI-MAP mean is given by

$$\mu_{jm} = \frac{\{\theta^{\rm num}_{jm}(\mathcal{O}) - \theta^{\rm den}_{jm}(\mathcal{O})\} + D_{jm}\hat{\mu}_{jm} + \tau^I \mu^{\rm map}_{jm}}{\{\gamma^{\rm num}_{jm} - \gamma^{\rm den}_{jm}\} + D_{jm} + \tau^I} \quad (16)$$

As with MMI training, this is an iterative process. At each stage the values of $\mu^{\rm map}_{jm}$ and $\sigma^{\rm map2}_{jm}$ are updated to reflect the changes in the numerator statistics.

The two free variables associated with MMI-MAP, τ and τ^I, have different effects. τ determines the center of the prior distribution for MMI-MAP. The smaller the value of τ the closer the prior distribution is to the ML model estimates. The value of τ^I determines the weight of the prior in the discriminative update. The larger τ^I is the closer the update will be to the prior distribution used. The value of τ^I is typically in the same range as used for I-smoothing (e.g. 100) and τ is normally in the range used for ML-MAP (e.g. 10).

3.3. MPE-MAP

In MPE [5], as for MMI, the auxiliary function to be optimised is represented in the form given in equation (11); but the statistics $\gamma^{\rm num}_{jm}$, $\gamma^{\rm den}_{jm}$ etc. are accumulated from the training data in a different way as described in [5]. The combination of the auxiliary function with the prior distribution used in I-smoothing follows the same pattern, with one difference: in MPE the numerator ("num") statistics are defined differently and do not correspond to the correct transcription. Therefore, where the correct-model statistics are needed (e.g., in equation 15) a separate set of statistics with the superscript "mle" are used in place of the "num" statistics; the "mle" statistics are the same statistics used in normal ML training.

4. Experiments

The performance of discriminative MAP was evaluated on two tasks. The first is to port a well-trained Switchboard system to the Voicemail task using limited training data. These results have previously been published in [6] and are summarised in this paper to allow an overview of the scheme. The second application examined is to build gender-specific HMMs using Broadcast News data by discriminative adaptation from gender independent models.

4.1. Porting Switchboard to Voicemail

Initial Switchboard HMMs were trained used 265 hours of data. Cross-word state-clustered triphones were generated. The system had 6684 distinct states and 16 Gaussians per state. For further details of the acoustic training see [1]. Two "initial" models were trained: an MLE-trained system and one discriminatively trained using MMIE. The Voicemail database consists of voicemail messages left by IBM employees. This data was partitions into a 94 minute test set and 28.1 hours of training data. The training data was further partitioned into nested subsets of approximately 1h, 4h, 15h and 20h. See [1] for more details of the database set-up.

All test set WERs reported here are from testing with a Switchboard language model (LMs). The baseline acoustic-model porting used a single iteration of ML-MAP. It was found that additional iterations yielded no further gains in performance. MMI-MAP task adaptation used four iterations of model parameter updates.The various forms of τ were approximately tuned, but there was little sensitivity to the precise values used.

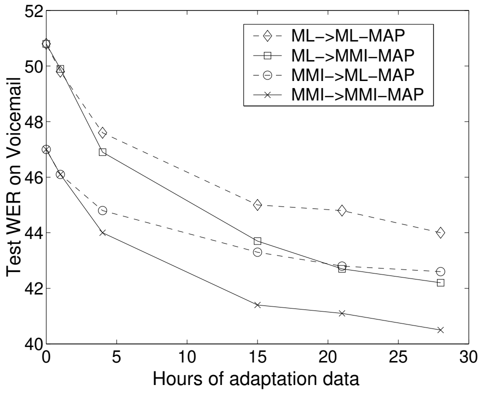

Figure 1: WERs for MMI-MAP and ML-MAP from MMI and ML baselines against amount of Voicemail adaptation data.

Figure 1 shows the word error rate (WER) when adapting either an ML or MMI-trained initial HMM set with ML-MAP or with MMI-MAP. The improvement from using an initial MMI-trained HMM set is retained if adaptation is with MMI-MAP but is partly lost with ML-MAP, especially with increasing amounts of adaptation data. There is 7.5% relative improvement from ML to MMI on the Switchboard-trained HMM set; the difference between ML-MAP-adapted ML and MMI-MAP-adapted MMI with 30h adaptation data is 8.0% relative. So the total improvement from discriminative training is 8.0%. Starting from the MMI-trained model, the improvement from using discriminative adaptation rather than ML adaptation is 4.6% relative.

4.2. Gender Dependent Broadcast News Models

The Broadcast News acoustic model training data consists of two sub-sets referred to as `BNtrain97` and `BNtrain98`, reflecting the years of their release. The combined set gives a total of 142 hours of training data [9]. A cross-word state-clustered triphone system was built using MLE with 6,976 speech states and 16 Gaussian components per state using MF-PLP parameterised speech with static, first and second order differences. MMIE and MPE trained models were also built. In addition a gender dependent system was generated using the training data speaker gender labels and only updating the Gaussian mixture weights and mean values. All experiments reported below used single pass decoding without adaptation. The decoder used a 65k word trigram language model which was taken from the 1998 Cambridge University broadcast news evaluation system [9]. The pronunciation dictionary was based on the 1993 LIMSI WSJ lexicon with many additions.

System	WER (%)	
	Std	HLDA
MLE-GI	19.6	17.9
MLE-GD	18.8	17.1
MMI-GI	17.0	—
MPE-GI	16.2	15.0
$\rightarrow$MPE-MAP	15.7	14.5

Table 1: WER on `BNeval98` using gender independent (GI) and gender dependent (GD) models with ML, MMI and MPE training and also MPE-MAP adaptation to GD models.

The error rates of the gender independent (GI) and gender dependent (GD) systems on the 1998 NIST Broadcast News evaluation data (`BNeval98`) is shown in table 1. Initially the system was tested using the standard front-end. The ML-GD system reduced the error rate by about 4% relative, 0.8% absolute, over the ML-GI system. Table 1 also shows the performance of MMI training and MPE training. Both discriminative training schemes show significant gains over ML training. MPE training gave a lower WER than MMI training yielding a 17% relative reduction in error rate over the MLE-GI system and 14% over the MLE-GD performance. As GD systems significantly reduced the error rate for the MLE system, it would be useful to generate gender dependent systems for the discriminative models. As the MPE-GI system outperformed the MMI-GI system, the MPE system was used as the original models for adaptation and MPE-MAP was applied. Table 1 lists the error rate for the MPE-GI system adapted with MPE-MAP to form GD models. These gender-dependent discriminative models gave an additional 3% relative reduction in WER over the MPE-GI system.

Table 1 also shows the performance of using the various training schemes with an HLDA frontend. Here third order differences were added to the feature vector and then projected down to 39 dimensions. The use of HLDA significantly reduced the WER for all systems. Using MPE-MAP yielded a 0.5% absolute reduction in error rate over the gender-independent system. An alternative approach to generating the GD model would rely on the I-smoothing to perform the regularisation and to simply do MPE training on the male and female training data separately. This gave an error rate of 14.8%, 0.3% higher than using MPE-MAP.

5. Conclusions

This paper has described techniques for incorporating prior information into discriminative training schemes. Versions based on both MPE, MPE-MAP, and MMI, MMI-MAP, have been described. It was shown that by using the appropriate form of the prior, these discriminative MAP schemes may be implemented by count smoothing. Depending on the exact form of the prior distribution used, this yields either versions of MAP estimation or I-smoothing. The discriminative adaptation schemes were investigated for both task porting, in this case from Switchboard to Voicemail, and for generating gender dependent models on the Broadcast News task. In both cases the methods were effective and allowed the performance advantage of discriminatively trained HMMs to be retained.

6. References

[1] M.J.F. Gales, Y. Dong, D. Povey & P.C. Woodland (2003). "Porting: Switchboard to the Voicemail Task", *Proc. ICASSP'03*, Hong Kong.

[2] J.L. Gauvain & C. Lee (1994). "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains." *IEEE Trans. SAP*, Vol. 2, pp. 291-299.

[3] Y. Normandin & S.D. Morgera (1991). "An Improved MMIE Training Algorithm for Speaker-Independent, Small Vocabulary, Continuous Speech Recognition", *Proc. ICASSP'91*.

[4] Y. Gao, B. Ramabhadran, M. Picheny (2000). "New Adaptation Techniques for Large Vocabulary Continuous Speech Recognition," *Proc. ICSA ITRW ASR2000*, Paris.

[5] D. Povey & P.C. Woodland (2002). "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," *Proc. ICASSP'02*, Orlando.

[6] D. Povey, P.C. Woodland & M.J.F. Gales (2003). "Discriminative MAP for Acoustic Model Adaptation," *Proc. ICASSP'03*, Hong Kong.

[7] L.F. Uebel & P.C. Woodland (2001). Discriminative Linear Transforms for Speaker Adaptation. *Proc. ISCA ITRW on Adaptation Methods for Automatic Speech Recognition*, Sophia-Antipolis.

[8] P.C. Woodland & D. Povey (2002). "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," *Computer Speech & Language* Vol. 16, pp. 25-48.

[9] P.C. Woodland (2002). The Development of the HTK Broadcast News Transcription System: An Overview, *Speech Communication*, Vol. 37, pp. 47-67.