

Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression

Asela Gunawardana and William Byrne

Center for Language and Speech Processing
Department of Electrical and Computer Engineering
The Johns Hopkins University,
3400 N. Charles St., Baltimore, MD 21218, USA
{zilla,byrne}@jhu.edu

Abstract

We present a simplified derivation of the extended Baum-Welch procedure, which shows that it can be used for Maximum Mutual Information (MMI) of a large class of continuous emission density hidden Markov models (HMMs). We use the extended Baum-Welch procedure for discriminative estimation of MLLR-type speaker adaptation transformations. The resulting adaptation procedure, termed Conditional Maximum Likelihood Linear Regression (CMLLR), is used successfully for supervised and unsupervised adaptation tasks on the Switchboard corpus, yielding an improvement over MLLR. The interaction of unsupervised CMLLR with segmental minimum Bayes risk lattice voting procedures is also explored, showing that the two procedures are complimentary.

1. Introduction

Discriminative training of acoustic hidden Markov models (HMMs) with Gaussian mixture emission densities [?] under a maximum mutual information (MMI) criterion has recently found to be useful in conversational large vocabulary continuous speech recognition (LVCSR) tasks [?]. In this paper we investigate the use of similar discriminative techniques in speaker adaptation using a variant of the popular maximum likelihood linear regression (MLLR) technique [?, ?]. The argument for estimating MLLR-style speaker adaptation transformations under a MMI criterion is very much the same as the argument for MMI estimation of speaker independent (SI) HMMs. While conventional maximum likelihood (ML) techniques would be optimal in the large data case if the true distribution of the acoustics was in fact an HMM, this not true in practice because the HMM assumption is highly unrealistic, and because we have finite training data in practice [?]. We present preliminary experiments on supervised and unsupervised speaker adaptation on the Switchboard corpus, which show that MMI estimates may in fact outperform ML estimates.

The paper is organized as follows. After a brief introduction to MMI estimation Section ??, we review the extended Baum-Welch procedure for MMI estimation of continuous emission density HMMs. We give a novel derivation of the extended Baum-Welch procedure, which shows that it can be used to find discriminative estimates of MLLR-type speaker adaptation transformations. In Sections ?? and ??, we present results using this algorithm for supervised and unsupervised adaptation on the Switchboard corpus. In Section ??, the interaction of this algorithm with a novel segmental minimum Bayes risk voting procedure [?] is explored.

2. MMI Estimation

In MMI estimation of acoustic HMMs [?, ?, ?], the goal is to find acoustic model parameters θ that maximize the mutual information estimate

$$\hat{I}(W_1^N, O_1^L; \theta) = \log \frac{q(\hat{w}_1^N, \hat{o}_1^L; \theta)}{q(\hat{w}_1^N) q(\hat{o}_1^L; \theta)}. \quad (1)$$

Here, the random variables W_1^N and O_1^L denote a word sequence of random length N and an acoustic observation vector sequence of random length L . The HMM state sequence S_1^L is hidden, and does not enter explicitly into the expression. The training data $(\hat{w}_1^N, \hat{o}_1^L)$ are a particular realization of (W_1^N, O_1^L) . Note that without loss of generality, we will represent the entire training (or adaptation) data by $(\hat{w}_1^N, \hat{o}_1^L)$, even when it consists of independent utterances. The language model probability $q_{W_1^N}(w_1^N)$ of a word sequence w_1^N and the acoustic model likelihood $q_{O_1^L|W_1^N}(o_1^L|w_1^N; \theta)$ of an acoustic observation sequence o_1^L given the word sequence w_1^N are denoted $q(w_1^N)$ and $q(o_1^L|w_1^N; \theta)$ for brevity, as is the marginal likelihood $q(o_1^L; \theta)$. It is easy to see that since the language model is independent of the acoustic model parameters θ , maximizing the mutual information estimate $\hat{I}(W_1^N, O_1^L; \theta)$ is equivalent to maximizing the conditional likelihood $q(\hat{w}_1^N|\hat{o}_1^L; \theta)$ of the training word sequence \hat{w}_1^N given the training acoustic sequence \hat{o}_1^L .

2.1. The Extended Baum-Welch Algorithm

Normandin [?] derives an extended Baum-Welch algorithm using a discrete approximation to the Gaussian density, and an extension of the Baum-Eagon inequality [?] due to Gopalakrishnan *et al* [?]. In the case where only the Gaussian means are updated (i.e. $\theta = \mu$), this derivation shows that the conditional likelihood is increased by reestimating the Gaussian means μ_s of the HMM states s according to

$$\mu_s^{(p+1)} = \frac{\sum_{\tau=1}^i (\gamma_s^{(p)}(\tau) - \gamma_s^{g(p)}(\tau)) \hat{o}_\tau + D \mu_s^{(p)}}{\sum_{\tau=1}^i (\gamma_s^{(p)}(\tau) - \gamma_s^{g(p)}(\tau)) + D}. \quad (2)$$

Here,

$$\gamma_s^{(p)}(\tau) = q_{S_\tau}(s|\hat{w}_1^N, \hat{o}_1^L; \theta^{(p)}) \quad (3)$$

is the conditional occupancy probability of state s at time τ given the training acoustics and transcription, and

$$\gamma_s^{g(p)}(\tau) = q_{s\tau}(s|\hat{o}_1^{\hat{i}}; \theta^{(p)}) \quad (4)$$

is the conditional occupancy probability of state s at time τ given the training acoustic data but not its transcription. Note that for simplicity we have assumed that the state emission densities are single Gaussians. There is no loss of generality in this, since HMMs with Gaussian mixture emission densities can be represented as HMMs with single Gaussian emission densities by using more states [?].

The extended Baum-Welch procedure can be derived without the discrete approximation to the Gaussian density mentioned above, and in fact applies to a general class of continuous HMM emission densities. We now present an outline of such a derivation which shows that the extended Baum-Welch procedure can be used for conditional maximum likelihood (CML) estimation of general continuous emission density HMMs. We apply it to the problem of estimating MLLR-type affine transformations of the Gaussian means. A detailed derivation is given in [?]. The derivation depends on the following two simple propositions, the first of which follows Gopalakrishnan *et al.* [?] and the second of which is a direct application of Jensen's inequality.

Proposition 1 Let $P(\theta) = \frac{Q(\theta)}{R(\theta)}$ be the ratio of two positive real valued functions Q and R on Θ . Given a value $\theta' \in \Theta$ and a real constant D' , define the function $F(\theta|\theta') = Q(\theta) - P(\theta')R(\theta) + D'$. Then $F(\theta|\theta') \geq F(\theta'|\theta')$ implies that $P(\theta) \geq P(\theta')$.

Proposition 2 Suppose f is a positive real valued function on $\mathcal{X} \times \Theta$, and $F(\theta) = \int_{\mathcal{X}} f(x, \theta) dx$. Then,

$$\int_{\mathcal{X}} f(x, \theta') \log f(x, \theta) dx \geq \int_{\mathcal{X}} f(x, \theta') \log f(x, \theta') dx$$

implies

$$F(\theta) \geq F(\theta').$$

Note that the integral over \mathcal{X} is a summation if \mathcal{X} is discrete.

We now observe that the conditional likelihood $q(\hat{w}_1^{\hat{n}}|\hat{o}_1^{\hat{i}}; \theta)$ is the ratio between the joint density $q(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{i}}; \theta)$ and the marginal density $q(\hat{o}_1^{\hat{i}}; \theta)$, so that we can apply Proposition ?? with any real constant D' and F defined as

$$F(\theta|\theta') = q(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{i}}; \theta) - q(\hat{w}_1^{\hat{n}}|\hat{o}_1^{\hat{i}}; \theta') q(\hat{o}_1^{\hat{i}}; \theta) + D'.$$

Taking \mathcal{X} to be the product space of all \hat{l} -length state sequences $s_1^{\hat{i}}$ and all \hat{l} -length observation vector sequences $o_1^{\hat{i}}$, we can write

$$F(\theta|\theta') = \sum_{s_1^{\hat{i}}} \int f(s_1^{\hat{i}}, o_1^{\hat{i}}; \theta | \theta') d o_1^{\hat{i}}, \quad (5)$$

with

$$\begin{aligned} f(s_1^{\hat{i}}, o_1^{\hat{i}}; \theta | \theta') &= [q(\hat{w}_1^{\hat{n}}, s_1^{\hat{i}}) \mathbf{1}_{\{\hat{o}_1^{\hat{i}}\}}(o_1^{\hat{i}}) \\ &\quad - q(\hat{w}_1^{\hat{n}}|\hat{o}_1^{\hat{i}}; \theta') q(s_1^{\hat{i}}) \mathbf{1}_{\{\hat{o}_1^{\hat{i}}\}}(o_1^{\hat{i}}) \\ &\quad + d'(s_1^{\hat{i}})] q(o_1^{\hat{i}}|s_1^{\hat{i}}; \theta), \quad (6) \end{aligned}$$

where $\mathbf{1}_{\{\hat{o}_1^{\hat{i}}\}}$ is the indicator function of the singleton set $\{\hat{o}_1^{\hat{i}}\}$, and $D' = \sum_{s_1^{\hat{i}}} d'(s_1^{\hat{i}})$.

Choosing d' (and therefore D') to ensure that f is positive, we can apply Proposition ?? with $F(\theta) = F(\theta|\theta')$. Thus choosing a parameter update $\theta^{(p+1)}$ so that

$$\sum_{s_1^{\hat{i}}} \int f(s_1^{\hat{i}}, o_1^{\hat{i}}; \theta^{(p)} | \theta^{(p)}) \log f(s_1^{\hat{i}}, o_1^{\hat{i}}; \theta | \theta^{(p)}) d o_1^{\hat{i}}$$

is maximized with respect to θ guarantees that $q(\hat{w}_1^{\hat{n}}|\hat{o}_1^{\hat{i}}; \theta^{(p+1)}) \geq q(\hat{w}_1^{\hat{n}}|\hat{o}_1^{\hat{i}}; \theta^{(p)})$. Dividing through by $q(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{i}}; \theta^{(p)})$ and using calculus to perform the maximization yields

$$\begin{aligned} \sum_{s_1^{\hat{i}}} \left\{ [q(s_1^{\hat{i}}|\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{i}}; \theta^{(p)}) - q(s_1^{\hat{i}}|\hat{o}_1^{\hat{i}}; \theta^{(p)})] \right. \\ \left. \cdot \nabla_{\theta} \log q(\hat{o}_1^{\hat{i}}|s_1^{\hat{i}}; \theta^{(p+1)}) \right. \\ \left. + d(s_1^{\hat{i}}) \int q(o_1^{\hat{i}}|s_1^{\hat{i}}; \theta^{(p)}) \nabla_{\theta} \log q(o_1^{\hat{i}}|s_1^{\hat{i}}; \theta^{(p+1)}) d o_1^{\hat{i}} \right\} \\ = 0, \quad (7) \end{aligned}$$

where $d(s_1^{\hat{i}}) = \frac{d'(s_1^{\hat{i}})}{q(\hat{w}_1^{\hat{n}}, \hat{o}_1^{\hat{i}}; \theta^{(p)})}$.

Substituting the Gaussian state emission density

$$q(o|s; \mu) = \frac{1}{\sqrt{2\pi}|\Sigma_s|} e^{-\frac{1}{2}(o-\mu_s)'\Sigma_s^{-1}(o-\mu_s)},$$

into equation (??) and choosing $d(s_1^{\hat{i}})$ appropriately yields the update rule of equation (??).

2.2. Conditional Maximum Likelihood Linear Regression

In MLLR [?, ?], the emission density of state s is reparametrized as

$$q(o|s; T) = \frac{1}{\sqrt{2\pi}|\Sigma_s|} e^{-\frac{1}{2}(o-T_r(s)\nu_s)'\Sigma_s^{-1}(o-T_r(s)\nu_s)},$$

with θ consisting of affine transformations T_r for regression classes $r = 1, \dots, R$. Substituting this into equation (??) can be shown to yield the following reestimation equation:

$$\begin{aligned} \sum_{s:r(s)=r} \left[\sum_{\tau=1}^{\hat{l}} (\gamma_s^{(p)}(\tau) - \gamma_s^{g(p)}(\tau)) + D_s \right] T_r^{(p+1)} \nu_s \nu_s' = \\ \sum_{s:r(s)=r} \left[\sum_{\tau=1}^{\hat{l}} (\gamma_s^{(p)}(\tau) - \gamma_s^{g(p)}(\tau)) \hat{o}_{\tau} + D_s T_r^{(p)} \nu_s \right] \nu_s', \end{aligned}$$

where $D_s = \sum_{s_1^{\hat{i}}} d(s_1^{\hat{i}}) \sum_{\tau=1}^{\hat{l}} \mathbf{1}_{\{s\}}(s_{\tau})$.

Since estimating speaker adaptation transformations using this reestimation equation increases the conditional likelihood $q(\hat{w}_1^{\hat{n}}|\hat{o}_1^{\hat{i}}; T)$, we call the resulting procedure *Conditional Maximum Likelihood Linear Regression* (CMLLR). Note that it differs from MLLR by the presence of the conditional occupancy probability $\gamma_s^{g(p)}$ given the acoustic data, and the relaxation term D_s .

In practice, $\gamma_s^{g(p)}(\tau)$ is found by performing the forward-backward procedure against $\hat{o}_1^{\hat{i}}$ using a recognition lattice

weighted by language model probabilities. This corresponds to the approximation

$$q(s_1^i | \hat{o}_1^i; \theta) \approx \frac{\sum_{w_1^n \in \mathcal{L}} q(w_1^n) q(s_1^i | w_1^n) q(\hat{o}_1^i | s_1^i; \theta)}{\sum_{w_1^n \in \mathcal{L}} \sum_{s_1^i} q(w_1^n) q(s_1^i | w_1^n) q(\hat{o}_1^i | s_1^i; \theta)}$$

The constants D_s , which control the ‘learning rate’ of the algorithm [?] are fixed to be $C \sum_{\tau=1}^i \gamma_s^{g(p)}(\tau)$ for some constant C , as suggested in [?]. In the following experiments, a unigram language model is used in computing $q(s_1^i | \hat{o}_1^i; \theta)$.

3. Supervised CMLLR Adaptation

Supervised adaptation experiments were performed on a subset of the the 2000 Hub-5 Switchboard-2 [?] evaluation set. A supervised adaptation task was simulated by splitting each conversation side in half and using one half as enrollment data (along with the true transcription) and the other half as test data, and vice versa. The test set was chosen so that the pronunciations of all enrollment words were known. This test set was composed of 866 utterances consisting of 10,260 words from 22 conversation sides, of 1 hour total duration.

The SI acoustic models were built using HTK [?] from 150 hours of Switchboard-1 and 14 hours of Callhome English data. The acoustic features used were 39-dimensional PLP cepstral coefficients with delta and acceleration components [?]. Cepstral mean and variance normalization, as well as speaker normalization through a bilinear transform [?] were performed over each conversation side. The acoustic models used cross-word triphones with decision tree clustered states [?], where questions about phonetic context as well as word boundaries were used in clustering. There were 8340 unique triphone states with 16 Gaussian components per speech state. Lattice rescoring experiments were performed using the AT&T Large Vocabulary Decoder [?], using a 33k-word trigram language model provided by SRI [?]. The SI word error rate (WER) is 38.7%.

WER results using MLLR and CMLLR on both the test and enrollment sets are shown in Figure ???. Both MLLR and CMLLR were performed with two regression classes, corresponding to speech and non-speech states, while the relaxation constant C was set to unity for CMLLR. It can be seen that the extended Baum-Welch algorithm is successful at improving the enrollment WER, as would be expected from a discriminative training algorithm. CMLLR alone gives an improvement (to 36.7% WER) over SI (38.7% WER), but fails to improve upon MLLR (36.3% WER). When CMLLR is used in addition to MLLR, a small improvement (to 35.8% WER) results.

4. Unsupervised CMLLR Adaptation

Unsupervised CMLLR adaptation experiments were performed in the course of system development for the 2001 Hub-5 LVCSR evaluation. Adaptation experiments were performed on the 1998 Hub-5 Switchboard-1 evaluation set (Swbd-1) [?] and the 2000 Hub-5 Switchboard-2 evaluation set (Swbd-2) [?]. The baseline system was as described above with the addition of speaker adaptive training (SAT) [?] and a refinement of the acoustic model state clustering.

The recognizer output from an intermediate stage of processing which did not use the refined state clustering was used as the reference transcription $\hat{w}_1^{\hat{n}}$ for adaptation; this transcription had a WER of 25.9% on Swbd-1 and 40.7% on Swbd-2. As shown in Table ??, CMLLR gives a small improvement over

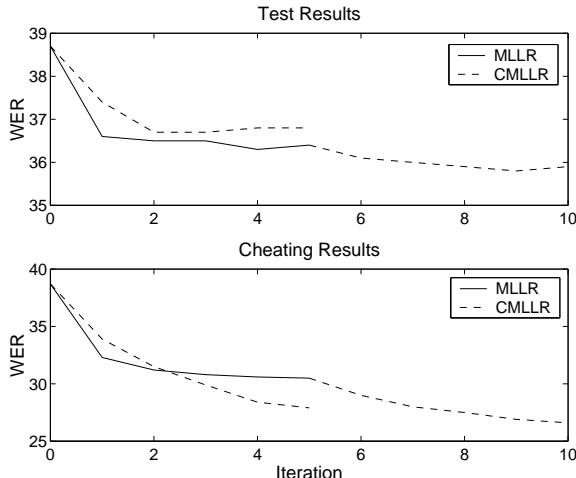


Figure 1: Comparison of WER on both the test and enrollment (cheating) sets for supervised adaptation using MLLR and CMLLR. Although CMLLR alone does not give as big an improvement over the SI WER as MLLR, it does provide a small improvement when used in addition to MLLR. Results of the cheating experiments show that CMLLR is more successful than MLLR in reducing the enrollment set WER.

	Swbd-1	Swbd-2
MLLR	25.7%	40.5%
MLLR + CMLLR	25.5%	40.2%
MLLR + LCER	25.0%	39.5%
MLLR + CMLLR + LCER	25.0%	39.5%

Table 1: Comparison of WER results for a first pass of unsupervised adaptation with MLLR and CMLLR on Swbd-1 and Swbd-2. MLLR indicates performing only MLLR and MLLR+CMLLR indicates performing CMLLR after MLLR. MLLR+LCER and MLLR+CMLLR+LCER indicate using lattices rescored with the corresponding adapted models for LCER.

MLLR on both test sets. However, when the lattice cutting e-ROVER (LCER) [?] segmental minimum Bayes risk voting procedure is used to combine hypotheses from the lattice, this gain is lost.

Since applying LCER to MLLR-rescored lattices makes a better transcription $\hat{w}_1^{\hat{n}}$ available, we perform a second pass of adaptation using this improved transcript. As shown in Table ??, the CMLLR results show a small improvement due to the better transcript, whereas the MLLR results do not. Furthermore, this gain is not lost when LCER is performed on the second pass.

5. Discussion

We have presented a simplified derivation of the extended Baum-Welch procedure, which shows that it can be used for CML estimation of arbitrary continuous emission density HMMs. We use this result to derive an algorithm to estimate MLLR-type speaker adaptation transformations of the HMM Gaussian means. Initial experiments on the Switchboard corpus show that the resulting CMLLR procedure is useful in both supervised and unsupervised adaptation tasks. However, these

	Swbd-1	Swbd-2
MLLR	25.5%	40.4%
MLLR + CMLLR	25.2%	39.9%
MLLR + LCER	24.9%	39.5%
MLLR + CMLLR + LCER	24.7%	39.2%

Table 2: Comparison of WER results for a second pass of unsupervised adaptation with MLLR and CMLLR on Swbd-1 and Swbd-2. The rows correspond to those of Table ??.

results are highly preliminary, and much remains to be done in determining the optimal manner in which to use this discriminative adaptation technique.

The question of how powerful a language model to use during estimation, and how it should be balanced relative to the acoustic model still need to be explored. Although the results above were obtained by performing the forward backward procedure over recognition lattices, the use of segmental procedures may actually perform better [?]. Using such procedure would allow the use of bigram and trigram language models during estimation. The unsupervised adaptation experiments show that CMLLR is more sensitive to the quality of the adaptation transcription than MLLR. Since CMLLR is a discriminative procedure, we would reasonably expect it to be more effective as the accuracy of the adaptation transcription improves. Here, we use LCER to obtain a better adaptation transcription for CMLLR, and believe that there may be deeper connections between the two procedures. Despite these open questions, the results presented here show that CMLLR behaves differently from MLLR, and that it can be used to improve recognition performance. How best to employ it is a focus of ongoing research.

6. Acknowledgements

We thank Andreas Stolcke for the use of the SRI language model and Michael Riley for the use of the AT&T Large Vocabulary Decoder. The lattice forward backward procedure was based on an implementation by Harriet Nock of Cambridge University, whose assistance is greatly appreciated. Many of the results presented here were obtained during the development of the CLSP 2001 LVCSR evaluation system. We are grateful to our colleagues Shankar Kumar and Veera Venkataramani who were part of this effort.

7. References

- [1] Y. Normandin, "Maximum mutual information estimation of hidden Markov models," in *Automatic Speech and Speaker Recognition: Advanced Topics* (C.-H. Lee, F. K. Soong, and K. K. Paliwal, eds.), ch. 3, pp. 57–81, Kluwer, 1996.
- [2] P. C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *Proc. ITRW ASR*, ISCA, 2000.
- [3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comp. Spch. & Lang.*, vol. 9, pp. 171–185, Apr. 1995.
- [4] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Spch. & Aud. Proc.*, vol. 3, pp. 357–366, Sept. 1995.
- [5] A. Nádas, D. Nahamoo, and M. A. Picheny, "On a model-robust training method for speech recognition," *IEEE Trans. Acoust., Spch., & Sig. Proc.*, vol. 36, pp. 1432–1436, Sept. 1988.
- [6] V. Goel, S. Kumar, and W. Byrne, "Confidence based lattice segmentation and minimum bayes-risk decoding of lattice segments," in *Eurospeech*, 2001. Submitted.
- [7] P. S. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Inf. Thry.*, vol. 37, pp. 107–113, Jan. 1991.
- [8] Y. Ephraim and L. R. Rabiner, "On the relations between modeling approaches for speech recognition," *IEEE Trans. Inf. Thry.*, vol. 36, pp. 372–380, Mar. 1990.
- [9] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information, and the Speech Recognition Problem*. PhD thesis, McGill University, Montreal, 1991.
- [10] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Am. Math. Soc.*, vol. 73, pp. 360–363, 1967.
- [11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, Version 3.0*, July 2000.
- [12] A. Gunawardana, "Maximum mutual information estimation of acoustic hmm emission densities," Tech. Rep. CLSP Research Note No. 40, CLSP, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA, 2001.
- [13] A. Martin, M. Przybocki, J. Fiscus, and D. Pallett, "The 2000 NIST evaluation for recognition of conversational speech over the telephone," in *Proc. Spch. Trans. Wkshp.*, NIST, 2000.
- [14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, Apr. 1990.
- [15] J. McDonough, W. Byrne, and X. Luo, "Speaker normalization with all-pass transforms," in *ICSLP*, 1998.
- [16] M. Mohri and M. Riley, "Integrated context-dependent networks in very large vocabulary speech recognition," in *Eurospeech*, 1999.
- [17] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. S. Onmez, F. Weng, and J. Zheng, "The SRI march 200 Hub-5 conversational speech transcription system," in *Proc. Spch. Trans. Wkshp.*, NIST, 2000.
- [18] A. Martin, J. Fiscus, M. Przybocki, and B. Fisher, "The evaluation: Word error rates and confidence analysis," in *Hub5 Wkshp.*, NIST, 1998.
- [19] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*, pp. 1137–1140, 1996.