SPEAKER-ADAPTED TRAINING ON THE SWITCHBOARD CORPUS

John McDonough *Tasos Anastasakos* *George Zavaliagkos* *Herbert Gish*

BBN Systems and Technologies
70 Fawcett Street, M/S 15/1c
Cambridge, MA 02138 USA
e-mail: *jmcd@bbn.com*

ABSTRACT

Speaker adaptation is the process of transforming some speaker-independent acoustic model in such a way as to more closely match the characteristics of a particular speaker. It has been shown by several researchers to be an effective means of improving the performance of large vocabulary continuous speech recognition systems. Until very recently speaker adaptation has been used exclusively as a part of the recognition process. This is undesireable inasmuch as it leads to a mismatched condition between test and training, and hence sub-optimal recognition performance. Very recently, there has been a growing interest in applying speaker-adaptation techniques to HMM training in order to alleviate the training/test mismatch. In prior work, we presented an iterative scheme for determining the maximum likelihood solution for the set of speaker-independent means and variances when speaker-dependent adaptation is performed during HMM training. In the present work, we shall investigate specific issues encountered in applying this general framework to the task of improving recognition performance on the Switchboard Corpus.

1. INTRODUCTION

Speaker adaptation is the process of transforming some speaker-independent (SI) acoustic model in such a way as to more closely match the characteristics of a particular speaker. Speaker adaptation has been shown by several researchers to be an effective means of improving the performance of large vocabulary continuous speech recognition (LVCSR) systems [3, 9, 8]. Several speaker adaptation paradigms have been proposed and investigated in the recent past. Until very recently, however, these schemes have been applied exclusively as a part of the recognition process. This is undesireable inasmuch as it leads to a mismatched condition between test and training, and hence sub-optimal recognition performance. Very recently, there has been a growing interest, both by the current authors and others, in applying speaker-adaptation techniques to HMM training; see, for example, Padmanabhan *et al* [7] and Anastasakos *et al* [1]. The latter works present an iterative scheme, dubbed speaker-adapted training (SAT), for determining the maximum likelihood (ML) solution for the set of speaker-independent means and variances when speaker-dependent (SD) adaptation is performed during HMM training. In the present work, we shall investigate specific issues encountered in applying this general framework to the task of improving LVCSR performance on the Switchboard Corpus.

The balance of this paper is organized as follows. In Section 2, we briefly review the iterative solution for the parameters of the SI model, which has appeared previously in [1]; this development will culminate with the statement maximum likelihood re-estimation formulae for the speaker-independent means and variances. Section 3 presents the results of several experiments conducted on the Switchboard Corpus; these experiments address issues pertaining to the optimal number of transformation parameters, the overall improvement in LVCSR performance achieved by SAT, the best way to initialize the SI model, and the effects of non-linear channels. Our conclusions and plans for future work are presented in Section 4.

2. SPEAKER-ADAPTED TRAINING

Here we give a brief review of the parameter optimization process inherent in SAT; a more complete treatment is available in [1, 2]. As with conventional HMM training, SAT begins with the formulation of an auxiliary function which must be maximized during the second stage of the expectation-maximization algorithm—see Dempster *et al* [5]. This function is conveniently expressed as

$$G(\mathcal{X}^{(s)}|A^{(s)},\Lambda) = \sum_{k,s} c_k^{(s)}[C_k - \tfrac{1}{2}(\tilde{\mu}_k^{(s)} - A^{(s)}\mu_k)^t D_k^{-1}(\tilde{\mu}_k^{(s)} - A^{(s)}\mu_k)] \quad (1)$$

where

μ_k and D_k are respectively the k^{th} SI mean and covariance matrix;

$\Lambda = \{\mu_k, D_k\}$ is the parameter set defining the SI model;

$c_k^{(s)} = \sum_i c_{ki}^{(s)}$ is the number of frames aligned to the k^{th} SI mean;

$x_i^{(s)}$is the i^{th} frame of speech data from speaker s;

$c_{ki}^{(s)}$ is the posterior probability that $x_i^{(s)} \sim \mathcal{N}(A^{(s)}\mu_k, D_k)$;

$\tilde{\mu}_k^{(s)} = \left(\sum_i c_{ki}^{(s)} x_i^{(s)}\right)/c_k^{(s)}$ is the k^{th} mean for speaker s;

$A^{(s)}$ is the speaker-dependent transformation matrix;

C_k is the logarithm of the Gaussian normalization constant.

In Eqn. (1) and what follows, we shall uniformly associate the index k with a specific Gaussian component and the index s with a given speaker. Our final objective is to jointly optimize three sets of parameters: the speaker-dependent transformation matrices $A^{(s)}$, the speaker-independent means μ_k and the speaker-independent variances D_k. Because of the coupling of these components present in (1), to do so directly would entail the simultaneous optimization of millions of parameters—an intractable problem. This quandary can be avoided, however, with the following stratagem: of the three sets of components, hold two of the same constant and solve, via a closed-form solution, for the optimal value of the third; iterate in this fashion over the three sets until the parameters converge to the desired optimum. A guarantee of convergence can be established by observing that the sub-optimization on each set of parameters must improve the global likelihood function, provided only that a fixed point has not been reached. In the limit of many iterations, the optimum value of the auxiliary function will be attained. Practical experience, however, indicates that a single iteration over each set of parameters between re-calculations of the auxiliary function yields a solution sufficiently close to the optimum; hence, in the sequel, we shall present a training paradigm based on this assumption.

The development necessary for determining the optimal transformation parameters has been published elsewhere: in the case of full or block-diagonal linear regression matrices the appropriate reference is [3]; for the case in which the components of the matrices are tied via an underlying conformal map see [9]; other formulations which result in a linear transformation are also possible [3, 8]. Thus, in what follows, we shall devote our attention exclusively to deriving the re-estimation formulae for the SI means and variances which comprise the heart of SAT, thereby illustrating the simplicity and intuitive appeal of this technique.

Mean Re-estimation

The mean re-estimation formulae are immediately available by taking the partial derivative of both sides of (1) with respect to the SI mean μ_k while holding all other terms fixed. The result can be expressed as

$$\mu_k = (A_k^t A_k)^{-1} A_k^t \tilde{\mu}_k \tag{2}$$

where

$$\begin{aligned} A_k^t A_k &= \sum_s c_k^{(s)} A^{(s)t} D_k^{-1} A^{(s)} && (3) \\ A_k^t \tilde{\mu}_k &= \sum_s c_k^{(s)} A^{(s)t} D_k^{-1} \tilde{\mu}_k^{(s)} && (4) \end{aligned}$$

Upon consideration of the solution in Eqns. (2–4), it is apparent that the estimation of the k^{th} ML speaker-independent mean is equivalent to the weighted least squares solution to a system of overdetermined linear equations—the solution has the form of the classical psuedo-inverse [4]. After some reflection, this comes as little surprise if we reason as follows: We can form the vector

$$\tilde{\mu}_k^t = \{\tilde{\mu}_k^{(1)t} \; \tilde{\mu}_k^{(2)t} \; \cdots \; \tilde{\mu}_k^{(S)t}\}$$

by concatentating all SD means and the matrix

$$A^t = \{A^{(1)t} \; A^{(2)t} \; \cdots \; A^{(S)t}\}$$

by concatenating all the SD transformation matrices. As is apparent from these definitions, both $\tilde{\mu}_k$ and the matrix-vector product $A\mu_k$ are elements of $\mathcal{R}^{(N \times S)}$, where N is the dimensionality of the feature vector and S the total number of speakers. The SI mean μ_k, however, is an element of $\mathcal{R}^N$. Hence, upon considering Eqn. (1) and ignoring the Gaussian normalization constant, it comes to light that we seek that μ_k achieving a minimum on the weighted Euclidean norm $||\tilde{\mu}_k - A\mu_k||$ where the weighting is determined by the co-variance matrix D_k. The optimal solution for μ_k is given by the perpendicular projection of $\tilde{\mu}_k$ onto the sub-space spanned by the columns of A—see Strang [4, Section 3.4].

Variance Re-estimation

A similarly straightforward set of equations can be developed for determining the optimal SI variances when all other components are fixed. Let us begin with an expression for the ML variance of the l^{th} dimension of the k^{th} Gaussian component

$$\sigma_{kl}^2 = \frac{1}{c_k} \sum_s \sum_i (x_{il}^{(s)} - \hat{\mu}_{kl}^{(s)})^2 c_{ki}^{(s)} \tag{5}$$

We now perform the following steps: add and subtract $\tilde{\mu}_{kl}^{(s)}$, the l^{th} component of the SD mean, to the quantity within parentheses; expand the square; cancel any zero terms; and define the SD variance

$$(\tilde{\sigma}_{kl}^{(s)})^2 = \frac{1}{c_k^{(s)}} \sum_i c_{ki}^{(s)} (x_{il}^{(s)} - \tilde{\mu}_{kl}^{(s)})^2 \tag{6}$$

This done, our final result is

$$\sigma_{kl}^2 = \frac{1}{c_k} \sum_s c_k^{(s)} \left[(\tilde{\sigma}_{kl}^{(s)})^2 + (\tilde{\mu}_{kl}^{(s)} - \hat{\mu}_{kl}^{(s)})^2 \right] \tag{7}$$

Training Paradigm

To illustrate the differences between conventional and speaker-adapted training, we present a schematic diagram of the latter in Figure 1. As is apparent from the figure, SAT extends the conventional training paradigm by requiring the re-estimation of SD adaptation parameters after each forward-backward pass. Subsequently, these parameters are used to re-estimate the SI means and variances as implied by Eqns (2–4) as well as (6–7). In addition, the SD adaptation parameters from the prior iteration are used to transform the SI codebook before the forward-backward stage. A final difference between SAT and conventional training stems from the fact that in the latter, the training utterances can be partitioned into any number of subsets suitable to expedite the training process; usually the number of such partitions is determined by the number of workstations available to perform the computation. In SAT, however, the training utterances must be partitioned according to speaker; hence one forward-backward pass with subsequent parameter re-estimation must be performed for every speaker in the set of training data. When training large models from corpora with many speakers, the necessity of partitioning the utterances by speaker can lead to a requirement on hard disk storage that is prohibitive. For this reasons, the actual algorithm used to implement the SAT re-estimation formulae is an area of current research.

In developing the re-estimation formulae above, the only assumption made about the transformations was that of linearity. To the knowledge of the authors, all adaptation paradigms currently being studied adhere to this assumption [3, 9, 8]. Hence, our framework for SI mean and variance re-estimation is broadly applicable.

3. APPLICATION TO SWITCHBOARD

The Switchboard Corpus is comprised of approximately 2,500 extemporaneous conversations between persons unknown to each other recorded over standard telephone lines. This speech itself is characterized by disfluencies of all sorts: restarts, filled pauses and partial words. The nature of the speech makes for a challenging problem when performing speech recognition in general or speaker adaptation in particular; chief among the issues that must be addressed in applying speaker adaptation to a corpus such as this are the need to learn the characteristics of a given speaker directly from the test data and the need to use errorful transcriptions obtained from a prior recognition pass in doing so. As we will discuss in greater detail below, non-linear channel effects can have a profound effect on speaker adaptation as well.

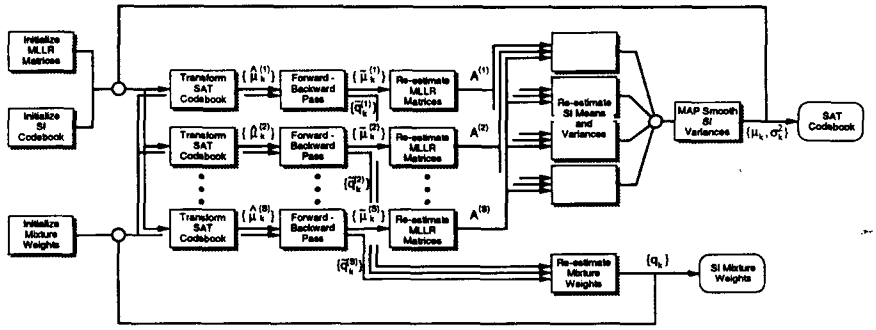

Figure 1. Speaker-adapted training schematic.

Adaptation	No. of Trans.	% WER
None	0	45.5
MLLR	1	43.2
MLLR	2	43.5
MLLR	4	43.5
MLLR	8	43.8
MLLR	16	44.3

Table 1. Word-error rate vs. number of MLLR transformations.

Model/Adaptation	% WER
Unadapted SI model	40.5
Adapted SI model	37.4
Adapted SAT model	35.6

Table 2. Improvement in word-error rate provdied by speaker-adapted training.

Number of Transformations

As the characteristics of a given speaker must be learned, not from some transcribed enrollment data, but from the test data itself, the number of transformation parameters which can reliably be estimated during speaker adaptation are fairly limited. Hence, prior to speaker-adapted training, it is necessary to establish an optimal operating point with regard to the number of transformations used—this can be accomplished most expeditiously be adapting the conventional SI with different numbers of transformations. Table 1 presents a sensitivity study of the number of transformations versus word error rate for a conventionally-trained phonetically-tied mixture system (PTM) with 256 Gaussian components per phoneme. The transformation in question is an MLLR matrix [3]. These results were obtained for recognition and adaptation on approximately five minutes of speech per speaker. As is apparent from the figure, a substantial improvement in accuracy is obtained from the use of a single transformation matrix; use of more transformation matrices only serves to reduce this initial gain. This is very much at odds with what is observed on corpora such as Wall Street Journal where there is either speaker-dependent enrollment data avaialable or the unadapted error rate is sufficiently low that several transformation matrices can be estimated; see Sankar *et al* [6].

Recognition Performance

The results in Table 2 illustrate the reduction in word error rate provided by speaker-adapted training. They were obtained using the state-clustered tied-mixture (SCTM) system used to run the 1996 Switchboard Evaluation, and illustrate that SAT is an effective means of reducing WER on conversational speech. In this case, the speaker-adapted training consisted of two EM passes, for each of which there was a single sub-iteration over the SD transformation parameters as well as the SI means and variances—see Figure 1. Speaker-adapted was initiated beginning with a conventionally-trained SI model. This same model was used for the unadapted decoding from which the errorful transcriptions necessary for unsupervised adaptation were obtained.

	SI Model	SAT Model	
		SI Seed	SD Seed
Unadapted	45.2	45.7	53.0
Adapted: 1st Pass	43.5	42.6	42.6
Adapted: 2nd Pass	N/A	42.0	41.8

Table 3. Recognition performance for SI model, SAT/SI model, and SAT/SD model, both adapted and unadapted.

Model Initialization

The goal of all study and investigation of speaker-dependent adaptation is the development of an SI system that, with minimal or no enrollment data, is able to achieve recognition performance comparable to an SD system trained on vastly more data. This observation and others following from it led us to speculate whether or not an SAT model trained from an SD seed might outperform a comparable model trained from an SI seed. Table 3 presents the results of our preliminary investigation in this area: we compare the performance of the SI model, the SAT model trained starting from the SI model (hereafter, SAT/SI), and the SAT model trained starting from the SD model (hereafter, SAT/SD). In all cases, adaptation is unsupervised and based on the errorful transcriptions from a previous unadapted decoding. All results are obtained from full four-pass decoding with PTM-256 models on the new Switchboard development test set. Based on these results, one might draw the following conclusions:

1. While the performance of the unadapted SAT/SI model degrades very little with respect to the unadapted SI model, the performance of the unadapted SAT/SD model is radically worse (ie, ~ 7.0% WER) than either.
2. After one alignment pass on the test data and subsequent adaptation, both SAT/SI and SAT/SD models give nearly identical performance. Both are approximately 1.0% better than the adapted SI model. Because it starts from much further back, however, this represents a greater gain for the SAT/SD model (~ 10.0%) than for the SAT/SI model (~ 3.0%).
3. After the second alignemnt pass on the test data and subsequent re-adaptation, the SAT/SD may have a miniscule advantage over the SAT/SI model.

The above would seem to indicate that, even though both the adapted SAT/SI and SAT/SD models perform very sim-

Side	Sup.	Uns.	No Adpt.	Deg.	Chnl.
2347-A	25.4	22.6	25.1	Y	
2347-B	31.6	35.0	38.6		
3469-A	50.7	41.7	43.6	Y	D
3469-B	49.9	46.5	50.1		
3520-A	58.5	65.1	72.5		
3520-B	30.6	28.9	30.4	Y	
3968-A	55.5	52.4	54.5	Y	D
3968-B	41.3	38.3	41.7		
4167-A	35.1	39.6	39.9		
4167-B	60.2	52.0	57.2	Y	
4622-A	58.7	57.8	60.9		
4622-B	46.9	40.2	41.8	Y	D
4771-A	37.8	38.9	37.3		
4771-B	56.3	56.3	60.7		
Ave.	45.0	43.5	45.2		

Table 4. Word error rate as a function of channel mismatch.

ilarly, they are actually quite different. This observation is given additional support by comparing the likelihoods of the respective unadapted models when initially aligned to the test data using the putative transcriptions from a prior decode; whereas the SAT/SI model generally exhibits a degradation in likelihood of 2-5% as compared to the SI model, the comparable figure for the SAT/SD model is 40-50%.

Channel Effects

An initial indication that channel effects may be significant on the Switchboard Corpus was obtained inadvertently when we performed several supervised adaptation experiments using the conventional SI PTM-256 model. Many Switchboard speakers participate in several conversations; this is true in particular of all speakers in our current development test set. The experiments in question were conducted by choosing a second conversation from among those held out of the training set for each speaker in the development test set, and using both it and its concomitant transcription as enrollment data for speaker-adaptation. Upon examining the recognition results, we were suprised to find the overall supervised adaptation performance to be virtually indistinguishable from that of the *unadapted* SI model—see Table 4. Moreover, it was worse than that of the adapted SI model where adaptation was unsupervised on the test conversation. In searching for an explanation for this anomaly, we broke down the recognition results by speaker; we found that for several speakers, supervised adaptation on the held-out conversation *degraded* performance, but unsupervised adaptation on the test conversation improved it. These conversation sides are marked with a "Y" in the "Degrade" column of Table 4. We then attempted to correlate these anonmalous sides with channel differences between the test conversation and the held-out conversation used for supervised adaptation; this was done by comparing the phone numbers associated with the respective conversations in found in table supplied with the Switchboard Corpus, a procedure that is less than conclusive as it does not account for the possibility of multiple hand sets sharing the same extension. Nonetheless, the results were fairly compelling: Those sides for which test and adaptation were performed on different channels (i.e., different numbers) are marked "D" in the "Channel" column. For each of these three conversation sides, WER degraded by at least 1.0%, substantially more in two of the three cases. This leaves three sides for which performance degraded when supervised adaptation was performed on the same channel as test; however, in all three cases the degradation was 1.0% or less, substantially less in two out of three.

These results are interesting inasmuch as for the purpose of SAT, we currently lump all training set (TS) conversations for a given speaker into one batch and then estimate a single set of adaptation parameters for the lot. Based on this very cursory analysis, it is apparent we should consider attempting to classify the conversations for a given speaker in terms of channel similarity, and grouping together only those that can plausibly be said to be the same. As a first step, we might hypothesize that the each conversation in the TS was collected across a different channel, and proceed accordingly.

4. CONCLUSIONS

Speaker-adapted training is a recent extension to the conventional EM-based training paradigm for continuous density hidden Markov models. In the limit, it provides a maximum likelihood solution for all parameters of the speaker-independent model when speaker-dependent adaptation is to be used on the test data. In this work, we have reported on our prelimary attempts to apply SAT to the conversational telephone-quality speech comprising the Switchboard Corpus. Beginning with an unadapted word error rate in the neighborhood of 40%, we found that an adapted SI model provides an improvement of $\sim$ 3.0% absolute over the unadapted model, while an adapted SAT model provides an additional improvement of $\sim$ 2.0%. We have also identified other issues that are candidates for future work, among these the possibility of obtaining speaker-dependent recognition performance by using an SD model to initiate SAT, and the need to compensate for non-linear channel effects introduced standard telephone handsets and lines.

REFERENCES

[1] Tasos Anastasakos, John McDonough, Richard Schwartz, John Makhoul, "A Compact Model for Speaker-Adaptive Training," to appear in International Conference on Spoken Language Processing (ICSLP) 1996.

[2] Tasos Anastasakos, John McDonough, John Makhoul, "Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization," submitted this proceedings.

[3] V.J. Leggetter and P.C. Woodland, "Speaker Adaptation Using Linear Regression", *Technical Report CUED/F-INFENG/TR.181*, Cambridge University, Engineering Department, June 1994.

[4] Strang, *Linear Algebra and Its Applications: Second Edition*, New York, Academic Press, 1980.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihoood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, B, 39, 1-38.

[6] A. Sankar, L. Neumeyer, and M. Weintraub, "An Experimental Study of Acoustic Adaptation Algorithms," *IEEE ICASSP*, pp. 713–716, May 1996.

[7] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulary Speech Recognition Systems," *IEEE ICASSP*, pp. 701–704, May 1996.

[8] G. Zavaliagkos, R. Schwartz and J. Makhoul, "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition", *IEEE ICASSP*, pp. 676–679, May 1995.

[9] J. McDonough, G. Zavaliagkos, and H. Gish, "An Approach to Speaker Adaptation Based on Analytic Functions," *IEEE ICASSP*, pp. 721-724, May 1996.