A COMPARATIVE STUDY OF SPEAKER ADAPTATION TECHNIQUES

Leonardo Neumeyer, Ananth Sankar and Vassilios Digalakis
e-mail: {leo,sankar,vas}@speech.sri.com

SRI International
Speech Technology and Research Laboratory
Menlo Park, CA, 94025, USA

ABSTRACT

In previous work, we showed how to constrain the estimation of continuous mixture-density hidden Markov models (HMMs) when the amount of adaptation data is small. We used maximum-likelihood (ML) transformation-based approaches and Bayesian techniques to achieve near native performance when testing nonnative speakers of the recognizer language. In this paper, we study various ML-based techniques and compare experimental results on data sets with recordings from nonnative and native speakers of American English. We divide the transformation-based techniques into two groups. In *feature-space* techniques, we hypothesize an underlying transformation in the feature-space that results in a transformation of the HMM parameters. In *model-space* techniques, we hypothesize a direct transformation of the HMM parameters. In the experimental section we show how the combination of the best ML and Bayesian adaptation techniques result in significant improvements in recognition accuracy. All the experiments were carried out with SRI's DECIPHER^{TM} speech recognition system [1][2].

1. INTRODUCTION

Automatic speech recognition (ASR) performance degrades rapidly when a mismatch exists between the training and the testing conditions. For example, the performance of ASR systems trained using native speakers degrades dramatically when tested on nonnative speakers [3]. Current methods to minimize the effect of such a mismatch include ML transformation-based approaches [4][5][6] and Bayesian adaptation [3][7][8].

This work focuses on comparing various ML transformation-based techniques and finding the optimum method for a given task. We also investigate combinations of these adaptation techniques.

2. THEORY

2.1. ML Adaptation Techniques

In ML transformation-based techniques [4][5][6], adaptation is achieved via a transformation of the speaker-independent observation densities. The transformation parameters θ_n are obtained by maximizing the likelihood of the adaptation data X given the corresponding word string W,

$$\theta_n = \text{argmax}_\theta p(X|\theta, W) \ . \quad (1)$$

A separate transformation is used for each group of Gaussian densities. The number of such transformations can be adjusted based on the available amount of adaptation data [4].

We assume that the speaker-independent (SI) HMM has state observation densities of the form

$$p_{SI}(y_t \mid s_t) = \sum_i p(\omega_i \mid s_t)\, N(y_t; \mu_{ig}, \Sigma_{ig}) \quad , \quad (2)$$

where g is the index of the Gaussian codebook used by state s_t. In this paper we investigate the adaptation of this system by jointly transforming all the Gaussians of each codebook.

As in [5], we consider transformations in two spaces: 1) the *feature-space*, and 2) the *model-space*. In the feature-space approach, the "original" features y_t are transformed to the observed features x_t by a hypothesized transformation $f_\nu(y_t)$ where ν are the parameters to be estimated. In the model-based approach, the original model λ_y is transformed to the new model λ_x by $\lambda_x = g_\eta(\lambda_y)$ where η are the parameters to be estimated.

The next two subsections describe the proposed transformation methods. Table 1 summarizes the methods.

2.1.1. Transformations in the Feature Space

Method I (Diagonal Affine Transform) [4]. In this method we assume that, given the HMM state index s_t, the observed features x_t can be obtained from the original features y_t through the transformation

$$x_t = A_g y_t + b_g \quad . \quad (3)$$

Under this assumption, the speaker-adapted (SA) observation densities will have the form

$$p_{SA}(x_t|s_t) = \sum_i p(\omega_i|s_t)\, N\left(x_t; A_g\mu_{ig} + b_g, A_g\Sigma_{ig}A_g^T\right) \quad (4)$$

the parameters $A_g, b_g, g = 1, \ldots, N_g$ are estimated using the ML approach of Eq. (1), where N_g is the number of distinct transformations. We use the EM algorithm [9] to derive the ML estimates of the parameters A_g, and b_g. When A_g is a diagonal matrix closed form estimates can be obtained for A_g and b_g as described in [4][5]. When A_g is a full matrix, however, the estimation problem is more tedious. In this paper we use a diagonal matrix.

Method II (Additive Transform). This case is identical to Method I with $A_g = I$.

Method III (Stochastic Additive Transform) [5]. In this case we model the acoustic mismatch using a stochastic transformation. The stochastic transform is given by,

$$x_t = y_t + b_g(\mu_{b,g}, \sigma_{b,g}) \quad , \tag{5}$$

where $b_g(\mu_{b,g}, \sigma_{b,g})$ is a Gaussian random variable with mean $\mu_{b,g}$ and variance $\sigma^2_{b,g}$. In this context, we can view Method II as a special case in which the additive term b_g is a deterministic parameter.

2.1.2. Transformations in the Model Space

Method IV (Full Affine Transform) [6]. An alternative to Method I is to transform the means of the Gaussian density functions using a full matrix and leave the variances unchanged. The advantage of using a full matrix is that we can model the correlation between feature components at the expense of a quadratic increase in the number of adaptation parameters. The observation densities in this case will have the form

$$p_{SA}(x_t|s_t) = \sum_i p(\omega_i|s_t)\, N(x_t; A_g\mu_{ig} + b_g, \Sigma_{ig}) \quad . \tag{6}$$

Method V (Structured Affine Transform). Our continuous density HMM system uses a single feature vector stream, which is the augmented vector composed of three basic feature vectors: cepstrum, first-derivative, and second-derivative of the cepstrum. The structured A_g matrix has non-zero values only in the elements whose rows and columns correspond to the same basic vector. For example, in Eq. (6) an element of the mean vector corresponding to a cepstrum component will only be predicted by the mean subvector that corresponds to the cepstrum and will not depend on the delta components.

The motivation for proposing this method is that the estimation of A_g involves inverting a sample correlation matrix. The dependencies between the cepstrum and its derivatives may result in an ill-conditioned sample correlation matrix, resulting in bad estimates.

Method VI (Scaled Variance Transform) [5]. In this case we transform the means using an additive shift and the variance using a scaling factor. The difference between this approach and Method I is that, in this case, the scale factor only affects the variance and is not tied to the scaling of the means.

2.2. Bayesian Adaptation Technique

In the Bayesian adaptation approach, the prior information is encapsulated in the SI models [3][7][8]. The Bayesian algorithms asymptotically converge to the speaker-dependent performance as the amount of adaptation speech increases. However, the adaptation rate is usually slow.

2.3. Combined Adaptation Technique

Finally, we use a combination of ML and Bayesian techniques to achieve the quick adaptation characteristics of the ML transformation-based methods with the asymptotic properties of Bayesian methods [3]. In this approach we first use the ML transformation-based method to adapt the SI models to the new speaker. These adapted models are then used as priors for the Bayesian adaptation method. The advantage of this approach is that the priors obtained by the ML transformation method are more closely matched to the observed data than the SI models.

Feature Space Methods			
Method	**Transf. Name**	**Mean Transform**	**Variance Transform**
I	Diagonal Affine	$\mu_i^{SA} = a_{ii}\mu_i^{SI} + b_i$	$\sigma^{2SA}_i = a^2_{ii}\sigma^{2SI}_i$
II	Additive	$\mu_i^{SA} = \mu_i^{SI} + b_i$	$\sigma^{2SA}_i = \sigma^{2SI}_i$
III	Stochastic Additive	$\mu_i^{SA} = \mu_i^{SI} + \mu_{b_i}$	$\sigma^{2SA}_i = \sigma^{2SI}_i + \sigma^2_{b_i}$
Model Space Methods			
Method	**Transf. Name**	**Mean Transform**	**Variance Transform**
IV	Full Affine	$\mu^{SA} = A_f\mu^{SI} + \boldsymbol{b}$	$\sigma^{2SA}_i = \sigma^{2SI}_i$
V	Structured Affine	$\mu^{SA} = A_s\mu^{SI} + \boldsymbol{b}$	$\sigma^{2SA}_i = \sigma^{2SI}_i$
VI	Scaled Variance	$\mu_i^{SA} = \mu_i^{SI} + b_i$	$\sigma^{2SA}_i = \alpha_i\sigma^{2SI}_i$

Table 1. Transformations of the means and variances for various methods.

3. EXPERIMENTS

Experiments were carried out using SRI's DECIPHER™ speech recognition system configured with a six-feature front end that outputs 12 cepstral coefficients, cepstral energy, and their first- and second-order derivatives. The cepstral features are computed from a fast Fourier transform (FFT) filterbank, and subsequent cepstral-mean normalization on a sentence basis is performed. We used genonic HMMs with an arbitrary degree of Gaussian sharing across different HMM states as described in [2]. The SI continuous HMM systems that we used as seed models for adaptation were gender-dependent, trained on 140 speakers and 17,000 sentences for each gender. Each of the two systems had 12,000 context-dependent phonetic models that shared 500 Gaussian codebooks (1000 in the native speaker experiments) with 32 Gaussian components per codebook. For testing, we used the Wall Street Journal (WSJ) task [10]. For fast experimentation, we used the progressive search framework [1]: an initial, SI recognizer with a bigram language model outputs word lattices for all the utterances in the test set. These word lattices are then rescored using SA models. We used the baseline WSJ 5,000-word (20,000 for native speaker experiments), closed-vocabulary bigram and trigram language models provided by the MIT Lincoln Laboratory. The trigram language model was used in the N-best rescoring paradigm, by rescoring the list of the N-best sentence hypotheses generated using the bigram language model.

3.1. Nonnative Speakers

We evaluated the adaptation algorithms on the 1994 "Spoke 3" task of the phase-1, large-vocabulary WSJ corpus [11][12]. For the first set of experiments we created a subset of the dev94 test set consisting of 5 nonnative speakers with 20 test sentences and 20 adaptation sentences per speaker. A bigram language model was used to compare performance between

the different adaptation methods. The experimental results are shown in Table 2. In the second column we indicate the adaptation method used (IV + III means we adapted using Method IV followed by Method III). All experiments were optimized for the number of transformations and only the best result is shown. The main conclusions from this experiments can be summarized as follows:

Non-Native Expts	Method	Number of Transforms	Word Error Rate (%)
NNat1	Speaker Independent	NA	24.9
NNat2	I	162	18.9
NNat3	I + Bayes	162	17.4
NNat4	II	162	19.2
NNat5	III	162	17.9
NNat6	VI	162	18.5
NNat7	IV	5	17.5
NNat8	IV + Bayes	5	14.4
NNat9	V	30	16.1
NNat10	IV + III	10/200	15.0
NNat11	IV + III + Bayes	10/200	15.2

Table 2. Word error rates for various supervised adaptation methods on a subset of the WSJ spoke3 (nonnatives) dev94 set using a bigram language model. Twenty adaptation sentences are used per speaker.

- Adapting the means with a Full transform produces better results (7% improvement) than adapting the means and the variances with a Diagonal transform (NNat2 vs NNat7).
- Bayesian adaptation helps when combined with ML adaptation, in both Diagonal and Full transform cases. In the diagonal case (NNat2 vs NNat3), we obtained an 8% improvement; in Full (NNat7 vs NNat8), we obtained an 18% improvement. Bayesian adaptation did not help in NNat11 when compared to NNat10.
- The Stochastic Additive transform is more effective than the deterministic Additive transform (NNat5 vs NNat4). In NNat10 and NNat11, the stochastic transform was used after the means were adapted with the Full Affine transform.
- The Structured Affine transform produced an improvement of 8% compared to the Full Affine case (NNat7 vs NNat9).

To see how this methods generalize when using a larger data set and more adaptation sentences, we used the best techniques on the full WSJ Spoke 3 development and evaluation sets. After some initial tests we decided to use Method IV followed by Method III followed by Bayesian adaptation. The results are presented in Table 3.These results show how the Full matrix transform and the Stochastic transform produced improvements of 20% in the dev94 set and 7% in the eval94 set over Method I + Bayes.

Data Set	SI	SA (I +Bayes)	SA (IV + III + Bayes)
S3 Dev 94	23.1	13.2	10.5
S3 Eval 94	23.2	11.3	10.5

Table 3. Speaker-independent and speaker-adapted word error rates on the WSJ Spoke 3 benchmark test using a trigram language model. Fourty adaptation sentences are used per speaker.)

3.2. Native Speakers

Some of the adaptation methods described in this paper were also tested on native speakers. We used 10 native speakers on the 20,000-word, closed vocabulary WSJ task (a total of 230 test sentences) and 40 adaptation sentences per speaker. The results are presented in Table 4. Unlike the nonnative case, we did not see a significant improvement after adapting the models using Method I (Nat2). The Full Affine transform (Nat3), however, produced a significant improvement of 14% after adaptation. Further improvement was gained when using the Structured Affine transform (Nat3 vs Nat5). The Bayesian adaptation produced some improvement in the Full case (Nat3 vs Nat4) and no significant improvement in the Structured case (Nat5 vs Nat6).

Native Expts	Method	Number of Transforms	Word Error Rate (%)
Nat1	Speaker Independent	NA	20.9
Nat2	I	160	20.5
Nat3	IV	2	17.9
Nat4	IV + Bayes	2	17.5
Nat5	V	10	17.6
Nat6	V + Bayes	10	17.5

Table 4. Word error rates for various supervised adaptation methods on natives speakers using a bigram language model. Forty adaptation sentences are used per speaker.

4. DISCUSSION

We compared six ML-based adaptation approaches and some combinations with Bayesian techniques. We found that transforming the means of the Gaussian density functions with a full matrix produces a significant improvement over the joint adaptation of the means and variances with a Diagonal transform. The variances can be adapted in a second stage using the Stochastic transform, and further improvement can be obtained in a third stage using Bayesian adaptation.

We also proposed a structured transformation of the means that overcomes the problem of inverting ill-conditioned sample correlation matrices. Other techniques, such as singular

value decomposition, can be used to overcome this problem and will be studied in more detail in the future.

Acknowledgments

This work was partially supported by ARPA through the Office of Naval Research Contract N00014-92-C-0154 and by Telia Research AB of Sweden.

The US Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Government funding agencies.

REFERENCES

1. H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER Speech Recognition System: Progressive Search Techniques," 1993 IEEE ICASSP, pp. II-319—II-322.
2. V. Digalakis and H. Murveit, "GENONES: Optimizing the Degree of Mixture Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer," 1994 IEEE ICASSP, pp. I537-I540.
3. V. Digalakis and L. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods," 1995 IEEE ICASSP, pp. I-680 - I-683.
4. V. Digalakis, D. Rtischev, and L. Neumeyer, "Fast Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," *IEEE Trans. on Speech and Audio Processing;* to appear.
5. A. Sankar and C.H. Lee, "Stochastic Matching for Robust Speech Recognition," IEEE Signal Processing Letters, Vol. 1, pp. 124-125, August 1994.
6. C.J. Leggetter and P.C. Woodland, "Flexible Speaker Adaptation using Maximum Likelihood Linear Regression," ARPA SLT Workshop, pp. 110-115, January 1995.
7. C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on Acoust., Speech and Signal Proc.,* Vol. ASSP-39(4), pp. 806—814, April 1991.
8. C.-H. Lee and J.-L. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters," 1993 IEEE ICASSP, pp. II-558 — II-561.
9. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood Estimation from Incomplete Data," *Journal of the Royal Statistical Society (B)*, Vol. 39, No. 1, pp. 1—38, 1977.
10. G. Doddington, "CSR Corpus Development," 1992 DARPA SLS Workshop, pp. 363-366.
11. D. Pallet, et al., "1994 Benchmark Tests for the ARPA Spoken Language Program," ARPA SLT Workshop, pp 5-36, January 1995.
12. F. Kubala, "Design of the 1994 CSR Benchmark Tests," ARPA SLT Workshop, pp 41-46, January 1995.